** Result [Patent] ** Format(P801) 01.Sep.2003 1/ 1

Application no/date: 2000-526867[1998/12/21]
Date of request for examination: [ ]
Public disclosure no/date: 2002-500393[2002/01/08] ¥ Translate
Examined publication no/date (old law): [ ]
Registration no/date: [ ]
Examined publication date (present law): [ ]
PCT application no PCT/US98/27199
PCT publication no/date WO/99/034291[1999/07/08]
Applicant: AVID TECHNOL INC
Inventor: PIITAAZU ERITSUKU SHII,RABINOUITSUTSU SUTANREE,JIEIKOBUSU HAABAAT O AARU,JIRETSUTO RICHIYAADO BEIKAA JIYUNIA,FUATSUSHIANO PIITAA JIEI
IPC: G06F 13/10 ,340 G06F 3/06 ,305 G06F 3/06 ,540 G06F 11/08 ,310 G06F 12/00 ,545 G06F 12/16 ,320 H04N 7/173 ,610
FI: G06F 11/08 ,310B G06F 13/10 ,340A G06F 12/16 ,320L G06F 3/06 ,305C G06F 3/06 ,540 G06F 12/00 ,545B H04N 7/173 ,610A
F-term: 5B001AA00,AB01,AC01,AD03,5B014EA04,EB04,FB03,FB04,GC07,GD23,GD32,GD33,HA09,5B018GA01,HA11,MA11,MA15,5B065BA01,EA03,EA12,EA19,EA24,EA31,EA35,ZA08,ZA15,5B082DE05,HA01,5C064BA07,BB06,BC18,BD02,BD13
Expanded classification: 452,446,451,453
Fixed keyword: R011,R012,R131
Citation:
Title of invention: The process which in scalable and dependability are high; and transfer hi
Abstract:

PURPOSE:Invocation, storage memory storing segment every each segment of chosen data are chosen, reading, series become in each segment of the data which called, transfer of independent high bandwidth data stream of high mass of dependability is enabled by supplying data in an application when data is received from the storage memory which syoteishi.

CONSTITUTION:Capture system makes segment table 90A.The pictorial image index which a map makes each pictorial image in off set to a data stream to take in is made.For example, the data which became index can support field or a frame.In addition,Capture system acquires list of available storage memory.Furthermore, when an application program to carry out on a client called for data, when each data segment is stored on at least two storage memory, claim seems to be satisfied, and storage memory is chosen.

(Machine Translation)

(19)日本国特許庁（ＪＰ）　(12)公表特許公報（Ａ）　(11)特許出願公表番号
特表2002－500393
（P2002－500393A）
(43)公表日　平成14年１月８日(2002.1.8)

(51)Int.Cl.[7]	識別記号	FI		テーマコード(参考)
G06F 13/10	340	G06F 13/10	340A	5B001
3/06	305	3/06	305C	5B014
	540		540	5B018
11/08	310	11/08	310B	5B065
12/00	545	12/00	545B	5B082

審査請求　未請求　予備審査請求　有　（全124頁）　最終頁に続く

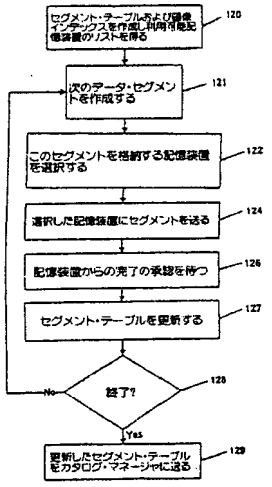
(21)出願番号　特願2000－526867(P2000－526867)
(86)(22)出願日　平成10年12月21日(1998.12.21)
(85)翻訳文提出日　平成12年６月26日(2000.6.26)
(86)国際出願番号　ＰＣＴ／ＵＳ９８／２７１９９
(87)国際公開番号　ＷＯ９９／３４２９１
(87)国際公開日　平成11年７月８日(1999.7.8)
(31)優先権主張番号　０８／９９７，７６９
(32)優先日　平成９年12月24日(1997.12.24)
(33)優先権主張国　米国（ＵＳ）
(31)優先権主張番号　０９／００６，０７０
(32)優先日　平成10年１月12日(1998.1.12)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　アヴィッド・テクノロジー・インコーポレーテッド
AVID TECHNOLOGY, INC.
アメリカ合衆国マサチューセッツ州01876, テュークスバリー, ワン・パーク・ウエスト, メトロポリタン・テクノロジー・パーク
(72)発明者　ピーターズ, エリック・シー
アメリカ合衆国マサチューセッツ州01741, カーライル, カールトン・ロード　80
(74)代理人　弁理士　社本　一夫　（外４名）

最終頁に続く

(54)【発明の名称】　コンピュータ・システムおよび多数の記憶装置および多数のアプリケーション間でスケーラブルにかつ信頼性高く多数の高帯域データ・ストリームを転送するプロセス

(57)【要約】
多数のアプリケーションは、コンピュータ・ネットワークを通じて多数の記憶装置からデータを要求する。データをセグメントに分割し、メディア・データの他のセグメントを格納した記憶装置には独立して、各セグメントを数個の記憶装置の１つにランダムに分散する。各セグメントに対応する冗長情報も、記憶装置間でランダムに分散する。セグメントに対する冗長情報は、セグメントのコピーとすることができ、各セグメントを少なくとも２つの記憶装置上に格納する。また、冗長情報は、２つ以上のセグメントに基づくことも可能である。このデータ・セグメントおよび対応する冗長情報のランダム分散により、スケーラビリティおよび信頼性双方の向上を図る。記憶装置が故障した場合、その負荷を均等に残りの記憶装置に分散し、その失われたデータを、冗長情報によって復元することができる。アプリケーションが選択したデータ・セグメントを要求した場合、最も短い要求キューを有する記憶装置がこの要求を処理することができる。多数のアプリケーションによって多数の記憶装置上に与えられる負荷のランダムな変動は、ほぼ等しく全



【特許請求の範囲】

【請求項１】　データを格納するための複数の記憶装置からなり、前記記憶装置上に格納したデータ・セグメントおよび対応する冗長情報を前記複数の記憶装置間でランダムに分散する、分散データ記憶システム。

【請求項２】　セグメントに対応する前記情報性情報は、当該セグメントのコピーである、請求項１記載の分散データ記憶システム。

【請求項３】　各セグメントの各コピーを前記記憶装置の異なるものに格納する請求項２記載の分散記憶システム。

【請求項４】　各セグメントの各コピーは、前記記憶装置の相対的仕様の関数として定義した確率分布にしたがって、前記複数の記憶装置の１つに割り当てられる請求項３記載の分散データ記憶システム。

【請求項５】　更に、コンピュータ読み取り可能ロジックを格納し、データ・セグメントの指示を用いてコンピュータによるアクセスが可能なセグメント・テーブルを定義するコンピュータ読み取り可能媒体を備え、前記セグメントおよび対応する冗長情報を格納した前記複数の記憶装置から、該記憶装置の指示を検索する請求項１記載の分散データ記憶システム。

【請求項６】　前記複数の記憶装置が、

コンピュータ・ネットワークに接続した第１記憶装置と、

前記コンピュータ・ネットワークに接続した第２記憶装置と、

前記コンピュータ・ネットワークに接続した第３記憶装置と、

を備える請求項１記載の分散データ記憶システム。

【請求項７】　セグメントに対応する前記冗長情報は、２つ以上のセグメントに基づく請求項１記載の分散データ記憶システム。

【請求項８】　前記２つ以上のセグメントおよび前記冗長情報を前記記憶装置の異なるものに格納する請求項７記載の分散データ記憶システム。

【請求項９】　コンピュータ用ファイル・システムであって、前記コンピュータが、当該コンピュータ上で実行するアプリケーションからの要求に応答してコンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上に格納してあるデータを読み出すことを可能とし、前記データの

セグメントおよび対応する冗長情報を前記複数の記憶装置間でランダムに分散するファイル・システムにおいて、

前記要求に応答してデータを読み出し、前記選択したデータの各セグメント毎に、前記セグメントを格納する記憶装置を特定する手段と、

前記要求したデータの各セグメントを、当該セグメントに対して特定した記憶装置から、読み出す手段と、

前記特定した記憶装置からデータを受信したとき、前記アプリケーションに前記データを供給する手段と、

を備えるファイル・システム。

【請求項１０】　セグメントに対応する前記冗長情報は前記セグメントのコピーであり、前記セグメントを記憶する前記記憶装置を特定する前記手段は、前記選択したデータの各セグメント毎に、当該セグメントを格納する記憶装置の１つを選択する手段を含む請求項９記載のファイル・システム。

【請求項１１】　前記記憶装置の１つを選択する手段は、前記複数の記憶装置上の要求の負荷が実質的に均衡するように記憶装置を選択することを特徴とする請求項１０記載のファイル・システム。

【請求項１２】　前記選択手段は、前記セグメントのためのどの記憶装置が、前記要求に応じるための推定時間が最も短いかについての推定に基づいて、前記セグメントのための前記記憶装置を選択する請求項１１記載のファイル・システム。

【請求項１３】　前記選択手段が、

前記ファイル・システムにおいて、

前記記憶装置の１つからデータを要求し、推定時間を示す手段と、

前記第１記憶装置が前記要求を拒絶した場合、前記記憶装置の他のものからデータを要求し、推定時間を示す手段と、

前記第２記憶装置が前記要求を拒絶した場合、前記第１記憶装置から前記データを要求する手段と、

を含み、

各記憶装置において、

前記要求が前記推定時間以内に前記記憶装置によって対応することができない場合、データの要求を拒絶する手段と、

前記要求が前記推定時間以内に前記記憶装置によって対応することができる場合、データの要求を受け入れる手段と、

を含む請求項１２記載のファイル・システム。

【請求項１４】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項１１記載のファイル・システム。

【請求項１５】　転送をスケジュールする前記手段は、

前記ファイル・システムにおいて、

前記選択した記憶装置から前記データの転送を要求し、待ち時間を示す手段と、

前記選択した記憶装置が前記データを転送する前記要求を拒絶した場合、別の記憶装置から前記データを要求する手段と、

を含み、

前記記憶装置において、

前記データが前記指示した待ち時間までに前記記憶装置から転送することができない場合、データ転送要求を拒絶する手段と、

前記選択した記憶装置が前記待ち時間以内に前記データを転送することができる場合、前記データを転送する手段と、

を含む請求項１４記載のファイル・システム。

【請求項１６】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項９記載のファイル・システム。

【請求項１７】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項７記載のファイル・システム。

【請求項１８】　転送をスケジュールする前記手段は、

前記ファイル・システムにおいて、

前記選択した記憶装置から前記データの転送を要求し、待ち時間を示す手段と、

前記選択した記憶装置が前記データを転送する前記要求を拒絶した場合、別の記憶装置から前記データを要求する手段と、

を含み、

前記記憶装置において、

前記データが前記指示した待ち時間までに前記記憶装置から転送することができない場合、データ転送要求を拒絶する手段と、

前記選択した記憶装置が前記待ち時間以内に前記データを転送することができる場合、前記データを転送する手段と、

を含む請求項１７記載のファイル・システム。

【請求項１９】　前記データを複数のセグメントに分割し、各セグメントをコピーし、各セグメントの各コピーを前記記憶装置の異なるものに格納する請求項１０記載のファイル・システム。

【請求項２０】　各セグメントの各コピーを、前記記憶装置の相対的仕様の関数として定義した確率分布にしたがって、前記複数の記憶装置の１つに割り当てる請求項１９記載のファイル・システム。

【請求項２１】　更に、コンピュータ読み取り可能ロジックを格納し、データ・セグメントの指示を用いてコンピュータによるアクセスが可能なセグメント・テーブルを定義するコンピュータ読み取り可能媒体を備え、前記セグメントおよび対応する冗長情報を格納した前記複数の記憶装置から、該記憶装置の指示を検索する請求項９記載のファイル・システム。

【請求項２２】　前記複数の記憶装置が、

前記コンピュータ・ネットワークに接続した第１記憶装置と、

前記コンピュータ・ネットワークに接続した第２記憶装置と、

前記コンピュータ・ネットワークに接続した第３記憶装置と、

を備える請求項９記載のファイル・システム。

【請求項２３】　セグメントに対応する前記冗長情報は、２つ以上のセグメントに基づく請求項９記載の分散データ記憶システム。

【請求項２４】　前記冗長情報および該冗長情報の基となる前記２つ以上のセグメントは、各々前記記憶装置の異なる１つに記憶されている請求項２３記載のファイル・システム。

【請求項２５】　コンピュータ用ファイル・システムであって、前記コンピュータが、当該コンピュータ上で実行するアプリケーションからの、要求に応答してコンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上にデータを格納することを可能とするファイル・システムにおいて、

前記データを格納する前記要求に応答して、前記データを複数のセグメントに分割する手段と、

各セグメントおよび該セグメントに対応する冗長情報を前記複数の記憶装置間でランダムに分散する手段と、

前記データを格納したか否かについて前記アプリケーションに確認する手段と

を備えるファイル・システム。

【請求項２６】　セグメントに対応する前記冗長情報は当該セグメントのコピーである請求項２５記載のファイル・システム。

【請求項２７】　前記ランダムに分散する手段は、

各セグメント毎に、ランダムにかつ他のセグメントのために選択した記憶装置とは独立して、少なくとも２つの前記記憶装置を選択する手段と、

前記選択した記憶手段に、各セグメントに対するデータを格納するように要求する手段と、

を備える請求項２６記載のファイル・システム。

【請求項２８】　前記選択する手段は、前記記憶装置の部分集合を選択する手段と、前記選択した部分集合内にある前記記憶装置の中から少なくとも２つの前記記憶装置を選択する手段とを含む請求項２７記載のファイル・システム。

【請求項２９】　各セグメントの各コピーは、前記記憶装置の相対的仕様の関数として定義した確率分布にしたがって、前記複数の記憶装置の１つに割り当てられる請求項２７記載のファイル・システム。

【請求項３０】　更に、コンピュータ読み取り可能ロジックを格納し、データ・セグメントの指示を用いてコンピュータによるアクセスが可能なセグメント・テーブルを定義するコンピュータ読み取り可能媒体を備え、前記セグメントおよび対応する冗長情報を格納した前記複数の記憶装置から、該記憶装置の指示を検索する請求項２５記載のファイル・システム。

【請求項３１】　前記複数の記憶装置が、

前記コンピュータ・ネットワークに接続した第１記憶装置と、

前記コンピュータ・ネットワークに接続した第２記憶装置と、

前記コンピュータ・ネットワークに接続した第３記憶装置と、

を備える請求項２５記載のファイル・システム。

【請求項３２】　セグメントに対応する前記冗長情報は、２つ以上のセグメントに基づく請求項２５記載の分散データ記憶システム。

【請求項３３】　前記冗長情報および該冗長情報の基となる前記２つ以上のセグメントは、各々前記記憶装置の異なる１つに記憶されている請求項３２記載のファイル・システム。

【請求項３４】　データを格納するための複数の記憶装置を備えた分散データ記憶システムにおいてデータを復元するプロセスであって、前記記憶装置上に格納した前記データのセグメントおよび対応する冗長情報は、前記複数の記憶装置間にランダムに分散してあり、前記プロセスは、前記記憶装置の１つの故障が検出されたときに実行され、

前記故障した記憶装置上に格納してあったコピーのセグメントを特定するステップと、

前記特定したセグメントに対応する前記冗長情報を格納した記憶装置を特定するステップと、

前記複数の記憶装置間において、前記特定した冗長情報に応じて前記特定したセグメントをランダムに分散するステップと、

からなるプロセス。

【請求項３５】　セグメントに対応する前記冗長情報は当該セグメントのコピーであり、

記憶装置を特定する前記ステップは、前記特定したセグメントの別のコピーがストレージであった記憶装置を特定するステップを含み、

ランダムに分散する前記ステップは、前記特定した記憶装置からの前記特定したコピーのコピーを前記複数の記憶装置間でランダムに分散するステップを含む、

請求項２４記載のプロセス。

【請求項３６】　セグメントに対応する前記冗長情報は、２つ以上のセグメントに基づき、更に、

前記特定したセグメントに対応する前記冗長情報にしたがって、前記特定したセグメントを再構成するステップを含み、

前記ランダムに分散するステップは、前記再構成した特定セグメントを前記複数の記憶装置間で分散するステップを含む、

請求項３４記載のプロセス。

【請求項３７】　ビデオ・データ・ストリームを組み合わせて複合ビデオ・データを生成し、ビデオ・データを格納するための複数の記憶装置を備えた分散システムに前記複合ビデオ・データを格納するプロセスであって、前記記憶装置上に格納した前記ビデオ・データのセグメントおよび対応する冗長情報を、前記複数の記憶装置間でランダムに分散する際に、

前記複数の記憶装置から前記ビデオ・データ・ストリームを読み出すステップと、

前記ビデオ・データ・ストリームを組み合わせ、前記複合ビデオ・データを生成するステップと、

前記複合ビデオ・データをセグメントに分割するステップと、

前記複合ビデオ・データのセグメントおよび対応する冗長情報を前記複数の記憶装置間でランダムに分散するステップと、

からなるプロセス。

【発明の詳細な説明】

【０００１】

（発明の分野）

本発明は、マルチメディア・プログラムのキャプチャ、オーサリングおよび再生のためのコンピュータ・システム、ならびに分散型計算機システムに関するものである。

【０００２】

（背景）

コンピュータ・ネットワークを通じてデータの分散使用に対応するいくつかのコンピュータ・システム・アーキテクチャがある。これらのコンピュータ・アーキテクチャは、社内イントラネット、分散データベース・アプリケーション、およびビデオ・オン・デマンド・サービスのような用途に用いられている。

【０００３】

例えば、ビデオ・オン・デマンド・サービスは、典型的に、ユーザがムービー全体を要求しており、選択されたムービーはかなりの長さを有するという仮定の下で設計する。したがって、ビデオ・オン・デマンド・サーバは、数人の加入者による同一ムービーへの、恐らくは異なる時点における、読み取り専用アクセスに対応するように設計されている。このようなサーバは、通常、データを数個のセグメントに分割し、これらのセグメントを数台のコンピュータまたはコンピュータ・ディスクに順次分散する。この技法は一般にストライピング（ｓｔｒｉｐｉｎｇ）と呼ばれており、例えば、米国特許第５，４７３，３６２号、第５，５８３，８６８号、および第５，６１０，８４１号に記載されている。数台のディスクにムービー用データをストライプする際の１つの問題として、１台のディスクまたはサーバの故障のために、ムービー全てが失われる可能性があることがあげられる。何故なら、あらゆるムービーは少なくとも１つのセグメントが各ディスク上に書き込まれているからである。

【０００４】

データ・ストレージにおける信頼性を高める一般的な技法に、ミラーリング（ｍｉｒｒｏｒｉｎｇ）と呼ばれるものがある。ミラーリングおよび順次ストライ

ピングを用いた混成システムが、米国特許第５，５５９，７６４号（Ｃｈｅｎ ｅｔ ａｌ．（チェンその他））に示されている。ミラーリングは、各記憶装置のコピーを２つ維持すること、即ち、全てのデータについて一次ストレージおよび二次バックアップ・ストレージを有する必要がある。コピーは双方とも、負荷分散のために使用することができる。しかしながら、この技法を用いると、一次ストレージが故障した場合、その負荷全体が二次バックアップ・ストレージ上にかかることになる。

【０００５】

数台のディスクにわたってデータを順次ストライプすることに伴う別の問題として、「コンボイ効果」（ｃｏｎｖｏｙ ｅｆｆｅｃｔ）と呼ばれるものの確度が高くなることがあげられる。コンボイ効果が発生するのは、あるファイルからのデータ・セグメントに対する要求が１つのディスクに集中し、次いでディスクからディスクに循環する（「コンボイ」）する傾向があるからである。その結果、１つのディスクに特定的に一度に要求の負担がかかり、一方他のディスクの負荷は軽くなる。また、ディスクに対する新しい要求は、そのいずれもがコンボイが処理されるのを待たなければならず、その結果、新たな要求に対するレイテンシが増大することになる。コンボイ効果を克服するためには、データをランダムにストライプすればよい。即ち、データ・ファイルのセグメントを、順次ではなく、ディスク間でランダムな順序で格納する。このようなシステムは、Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ Ｍｕｌｔｉｍｅｄｉａ’９６、ｐｐ１４４～１５０に、Ｒ．Ｔｅｗａｒｉ ｅｔ ａｌ．（Ｒ．テワリその他）による”Ｄｅｓｉｇｎ ａｎｄ Ｐｅｒｆｏｒｍａｎｃｅ Ｔｒａｄｅｏｆｆｓ ｉｎ Ｃｌｕｓｔｅｒｅｄ Ｖｉｄｅｏ Ｓｅｒｖｅｒｓ”（クラスタ化ビデオ・サーバにおける設計および性能のトレードオフ）に記載されている。このようなシステムでもなお、１つのディスク上でのランダムな過剰負荷が生ずるが、これは一般的なデータ・アクセスのランダム性によるものである。

【０００６】

これらのシステムには、多数の記憶装置および多数のアプリケーション間においてスケーラブルにかつ信頼性高く、多数の独立した高帯域データ・ストリーム

、特に、ビデオおよび関連するオーディオ・データのような等時性メディア・データを個別に転送可能なものはない。このようなデータ転送の必要性は、特に、マルチメディア・データのキャプチャ、オーサリング、および再生に対応するシステムでは困難である。特に、オーサリング・システムでは、データへのアクセスは、典型的に、より大きなデータ・ファイルの、クリップと呼ばれる、小断片を単位として行われる。これらのクリップは、どのようにデータを格納するかに関しては、任意に即ちランダムな順序でアクセスする場合が多く、効率的なデータ転送を行なうことが困難である。

（概要）

コンピュータ・ネットワークを用いて、多数のアプリケーションと接続してある多数の記憶装置上にデータをランダムに分散する。データをセグメントに分割する。各セグメントを記憶装置の１つに格納する。また、１つ以上のセグメントに基づく冗長情報も、基本となるセグメントとは異なる記憶装置上に格納する。冗長情報は、各セグメントのコピーとすればよく、あるいは２つ以上のセグメントに対して行なう排他的ＯＲ演算によって計算してもよい。セグメントまたは冗長情報を格納した各記憶装置の選択はランダムまたは疑似ランダムであり、他のデータ・セグメントが格納されている記憶装置とは独立することができる。冗長情報が２つ以上のセグメントに基づく場合、セグメントの各々を異なる記憶装置上に格納する。

【０００７】

このデータ・セグメントのランダム分散によって、スケーラビリティ（ｓｃａｌａｂｉｌｉｔｙ）および信頼性双方が向上する。例えば、セグメントにアクセスすることによってデータを処理するので、データ・フラグメント即ちクリップもデータの全てと同様に効率的に処理される。アプリケーションは、データ転送が効率的な場合にのみ、記憶装置からのデータ転送を要求することができ、更に記憶装置にリード要求を前処理するように要求することも可能である。コンピュータ・ネットワーク上での帯域幅利用は、クライアントおよび記憶装置間におけるデータ転送をスケジューリングすることによって最適化することができる。記憶装置の１つが故障した場合、その負荷も残りの記憶装置全体にランダムにそし

てほぼ均一に分散される。記憶装置の故障から復元する手順も備えることができる。

【０００８】

記憶装置およびアプリケーションは、中央制御部を用いずに、独立して動作することも可能である。例えば、各クライアントは、ローカル情報のみを用いて、記憶装置との通信をスケジュールすることができる。したがって、記憶装置およびアプリケーションのシステムに対する追加および除去が可能となる。その結果、システムは動作中でも拡張可能となる。

【０００９】

冗長情報が１セグメントのコピーである場合、システム性能を向上させることができるが、記憶増大が犠牲となる。例えば、アプリケーションが選択したデータ・セグメントを要求する場合、要求は、要求のキューが最も短い記憶装置によって処理することができるので、多数の記憶装置上の多数のアプリケーションによって加えられる負荷におけるランダムな変動は統計的にそして一層等しく記憶装置全てにわたって均衡化される。

【００１０】

この技法の組み合わせによって、多数の記憶装置および多数のアプリケーション間でスケーラブルにかつ信頼性高く多数の独立した高帯域幅データ・ストリームを転送することができるシステムが得られる。

【００１１】

したがって、一態様では、分散データ記憶システムは、データを格納する複数の記憶装置を含み、記憶装置上に格納したデータのセグメントを、複数の記憶装置間で分散する。各セグメントに対応する冗長情報も、記憶装置間にランダムに分散する。

【００１２】

冗長データが１セグメントのコピーである場合、各セグメントの各コピーは、記憶装置の異なるものに格納することができる。各セグメントの各コピーは、記憶装置の相対的指定（ｒｅｌａｔｉｖｅ　ｓｐｅｃｉｆｉｃａｔｉｏｎ）の関数として定義した確率分布にしたがって、複数の記憶装置の１つに割り当てること

ができる。分散データ記憶システムは、コンピュータ読み取り可能媒体を含むことができる。この媒体上にはコンピュータ読み取り可能なロジックを格納してあり、データ・セグメントの指示を用いてコンピュータによってアクセス可能なセグメント・テーブルを定義し、セグメントのコピーを格納してある複数の記憶装置から、記憶装置の指示を検索する。複数の記憶装置は、コンピュータ・ネットワークに接続してある、第１、第２および第３記憶装置を含むことができる。

【００１３】

別の態様では、コンピュータ用ファイル・システムは、コンピュータ上で実行するアプリケーションからの、要求に応答してコンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上に格納してあるデータを読み出すことを可能とする。前記データのセグメントおよび対応する冗長情報を前記複数の記憶装置間でランダムに分散する。冗長情報がセグメントのコピーである場合、ファイル・システムは、要求に応答してデータを読み出し、選択したデータの各セグメント毎に、セグメントを格納する記憶装置の１つを選択する。ファイル・システムは、他のセグメントおよび冗長情報から、失われたセグメントを再構成することができる。要求データの各セグメントは、当該セグメントのために選択した記憶装置から読み出される。選択した記憶装置からデータを受信すると、データをアプリケーションに供給する。このファイル・システムでは、複数の記憶装置上の要求の負荷が実質的に均衡化するように、記憶装置を選択することができる。セグメントのための記憶装置は、要求に対応する推定時間が最も短い、セグメントのための記憶装置の推定値に応じて選択することができる。

【００１４】

更に特定すれば、ファイル・システムは、記憶装置の１つからデータを要求し、推定時間を示すことができる。第１記憶装置が要求を拒絶した場合、ファイル・システムは別の記憶装置からデータを要求することができ、別の推定時間を示すことができる。第２記憶装置が要求を拒絶した場合、ファイル・システムは第１記憶装置からデータを要求する。各記憶装置は、推定時間以内に当該記憶装置が要求に応じられないときに、データ要求を拒絶する。記憶装置は、推定時間以

内に当該記憶装置が要求に応じられるときに、要求を受け入れる。

【００１５】

ファイル・システムは、選択した記憶装置からのデータ転送をスケジュールし、記憶装置が効率的にデータを転送するように、各セグメントを読み出すことができる。更に特定すれば、ファイル・システムは、選択した記憶装置からのデータ転送を要求し、待ち時間を示すことができる。選択した記憶装置がデータを転送する要求を拒絶した場合、他の記憶装置からデータを要求することができ、またはファイル・システムが後の時点で同じ記憶装置からデータを要求することも可能である。各記憶装置は、示された待ち時間以内に記憶装置からデータを転送することができない場合、データ転送の要求を拒絶する。記憶装置は、示された待ち時間以内に、選択した記憶装置がデータを転送することができる場合、データを転送する。

【００１６】

別の態様では、コンピュータ用ファイル・システムは、コンピュータ上で実行するアプリケーションからの要求に応答して、コンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上にデータを格納することを可能とする。ファイル・システムは、データ格納要求に応答して、データを複数のセグメントに分割する。各セグメントは、１つ以上のセグメントに基づく冗長情報と共に、複数の記憶装置間でランダムに分散される。ファイル・システムは、アプリケーションに、データを格納するか否か確認する。

【００１７】

このファイル・システムでは、冗長データがセグメントのコピーである場合、データのランダム分散は、各セグメント毎に、ランダムにかつ他のセグメントに選択した記憶装置には独立して、少なくとも２つの記憶装置を選択することによって行なうことができる。選択した記憶装置に、各セグメント毎にデータを格納するように要求することができる。ファイル・システムは、記憶装置の部分集合を選択することができ、選択した部分集合内にある記憶装置から、セグメントを格納する記憶装置を選択することができる。

【００１８】

また、ファイル・システムの機能性は、他のアプリケーションによって、またはアプリケーション・プログラム・インターフェースを通じてアクセス可能なコード・ライブラリによっても得ることができる。したがって、別の態様は、記憶装置の選択およびネットワーク転送のスケジューリングを含む、リードまたはライト機能を実行するように実現したクライアントまたはプロセスである。別の態様は、記憶装置の選択およびネットワーク転送のスケジューリングを含むリードまたはライト機能を実行するために、これによって実現した記憶装置またはプロセスである。別の態様は、このような機能性を実現する分散コンピュータ・システムである。これらの動作は、クライアントまたは記憶装置によって、ローカル情報のみを用いて実行することができ、システムを容易に拡張可能とする。

【００１９】

別の態様では、データを記憶するための複数の記憶装置を有する分散データ記憶システムにおいてデータを復元する。記憶装置の１つの故障が検出された場合、記憶装置上に格納したデータ・セグメントおよび冗長情報を複数の記憶装置間にランダムに分散する。データを復元するために、故障した記憶装置上にコピーを格納したセグメントを特定する。特定したセグメントに対応する冗長データを格納した記憶装置を特定する。冗長情報を用いて、特定したセグメントのコピーを再構成し、次いでこれらを複数の記憶装置間でランダムに分散する。このようなデータ復元は、ここに記載するファイル・システムまたは分散記憶システムのリードおよびライト機能性と組み合わせて用いることができる。

【００２０】

別の態様では、ビデオ・データ・ストリームを組み合わせて複合ビデオ・データを生成し、ビデオ・データを格納するための複数の記憶装置を備えた分散システムに格納する。記憶装置上に格納したビデオ・データのセグメントのコピーを、複数の記憶装置間でランダムに分散する。ビデオ・データ・ストリームを含むの記憶装置から読み出す。これらのビデオ・データ・ストリームを組み合わせて、複合ビデオ・データを生成する。複合ビデオ・データをセグメントに分割する。複合ビデオ・データのセグメントのコピーを、複数の記憶装置間でランダムに分散する。データの読み取りおよび格納は、ここに記載する技法を用いて行なう

ことができる。

（詳細な説明）

添付図面に関連付けて読むべき以下の詳細な説明では、本発明の実施形態の例を明示する。ここに引用する全ての引例は、ここで言及することによって明示的に本願にも含まれるものとする。

【００２１】

多数のアプリケーションおよび多数の記憶装置間において、データ転送、特に、モーション・ビデオおよび付随するオーディオ、ならびにその他の時間的に連続するメディアのような、多数の独立した高帯域幅時間厳守データ・ストリームの転送に対応するスケーラブルで信頼性の高い分散システムの設計において、いくつかの問題点が生ずる。このようなシステムでは、例えば、モーション・ビデオ・プログラムのオーサリングを行なうために用いるアプリケーションは、数個の記憶装置にわたって分散している可能性がある数個の異なるファイルの数個の小部分にランダムにアクセスする場合がある。数個のアプリケーションが同じデータに即時かつ同時のアクセスを要求する場合もあり、いずれのアプリケーションもいずれの時点でもいずれのメディア・ピースにもアクセスすることができなければならない。放送またはその他時間厳守再生に用いるシステムでは、フォールト・トレランスも望ましい。最後に、システムは、新たな記憶装置や新たなアプリケーションの追加を、システムの動作中においても簡単にできるように、拡張可能かつスケーラブルでなければならない。このようなシステムに望ましいその他の特性には、故障までの平均時間が長いこと、単一点で故障しないこと、迅速にかつ動作しながらでも修理可能であること、動作を中断することなく記憶装置の故障に対する耐性があること、そして失われたデータを復元可能であることが含まれる。

【００２２】

一実施形態において、本システムは、コンピュータ・ネットワークによって多数の別個で独立したデータ記憶用記憶装置に接続した多数のアプリケーションを含む。データはセグメントに分割する。各セグメント毎の冗長情報を決定し、セグメントおよびその冗長情報を別々の記憶装置上に格納する。セグメントに対す

る記憶装置の選択はランダムまたは疑似ランダムであり、直前のセグメントのような、他のセグメントに選択した記憶装置とは独立とすることもできる。冗長情報およびランダムなデータ分散双方によって、アプリケーションおよびストレージ間において双方向にデータを効率的に転送するシステムの機能が向上し、フォールト・トレランス性が高くなる。

【００２３】

冗長情報は、セグメントのコピーとすることができる。このセグメントの複製により、システムは、要求のキューが最も短い記憶装置を選択することによる等のように、特定のアプリケーションがどの記憶装置をアクセスするのかも制御することが可能となる。その結果、負荷のランダムな変動は、記憶装置全てにほぼ均等に分散される。

【００２４】

また、アプリケーションは、転送が効率的な場合にのみ、記憶装置にデータ転送を要求することも可能である。ネットワーク上での通信を適切にスケジュールすることにより、ネットワークの輻輳を低減することができ、ネットワークの帯域幅を一掃効率的に使用することができる。各クライアントにローカル情報を使用して記憶装置との通信をスケジュールさせることによって、中央制御点を不要とすることも可能である。

【００２５】

図１Ａは、一例としてのコンピュータ・システム４０を示す。このコンピュータ・システムは複数の記憶装置４２を含む。記憶装置とは、ディスクのような不揮発性コンピュータ読み取り可能媒体を備え、その上にデータを格納することができるデバイスのことである。また、記憶装置は、高速な、典型的に不揮発性のメモリも有し、その中に媒体からのデータを読み出す。各記憶装置は、それ自体の独立したコントローラも有し、リードおよびライト・アクセスを含みこれらには限定されない、媒体上に格納してあるデータへのアクセス要求に応答する。例えば、記憶装置４２は、サーバ・コンピュータとすることができ、サーバのファイル・システム内にあるデータ・ファイルにデータを格納する。コンピュータ・システム４０には任意の数の記憶装置があり得る。

【００２６】

アプリケーション４４は、コンピュータ・ネットワーク４６を通じて記憶装置への要求を介して、記憶装置４２へのアクセスを要求するシステムである。記憶装置４２は、コンピュータ・ネットワーク４６を通じて、アプリケーション４４にデータを送出したり、あるいはアプリケーション４４からのデータを受け取る。アプリケーション４４は、ディジタルまたはアナログ・ソースから受け取ったデータを捕獲し、記憶装置４２上にデータを格納するシステムを含むことができる。また、アプリケーション４４は、マルチメディア・プログラムのオーサリング、処理または再生用システムのように、記憶装置からデータを読み出すシステムも含むことができる。他のアプリケーション４４は、種々の故障回復タスクを実行することができる。また、アプリケーション４４を「クライアント」と呼ぶ場合もある。１つ以上のカタログ・マネージャ４９も使用可能である。カタログ・マネージャとは、アプリケーション４４によるアクセスが可能なデータベースであり、記憶装置４２上で利用可能なデータに関する情報を保持している。この実施形態は、１９９７年１０月２３日付けのＰＣＴ公開ＷＯ９７／３９４１１に示されるような放送ニュース・システムを実現するために用いることも可能である。

【００２７】

記憶装置４２に格納するデータは、セグメントに分割する。１つ以上のセグメントに基づいて、冗長情報を作成する。例えば、各セグメントをコピーすればよい。その結果、少なくとも２つの記憶装置４２上に格納各セグメントを格納することになる。あるいは、２つ以上のセグメントの排他的ＯＲによって、冗長情報を作成することも可能である。各セグメントは、その冗長情報とは異なる記憶装置４２に格納する。セグメントおよびその冗長情報を格納する記憶装置の選択はランダムまたは疑似ランダムであり、他のデータ・セグメントを格納する記憶装置からは独立とすることも可能である。一実施形態では、同じ記憶装置には、２つの連続するセグメントを格納しない。セグメントおよびその冗長情報を格納する記憶装置を選択するための確率分布は、記憶装置全体にわたって均一とするとよく、容量、帯域幅、およびレイテンシというような、記憶装置の仕様は同様と

する。この確率分布は、各記憶装置の仕様の関数とすることも可能である。データ・セグメントおよび対応する冗長情報のランダム分散により、スケーラビリティおよび信頼性双方が向上する。

【００２８】

データ・セグメントのコピーのランダム分散の一例を図１Ａに示す。図１Ａにおいて、ｗ，ｘ，ｙおよびｚで示す４つの記憶装置４２が、１，２，３および４で示す４つのセグメントに分割したデータを格納する。セグメントおよびそのコピーのランダム分散の一例を示す。ここで、セグメント１および３を記憶装置ｗに格納し、セグメント３および２を記憶装置ｘに格納し、セグメント４および１を記憶装置ｙに格納し、セグメント２および４を記憶装置ｚに格納する。

【００２９】

図１Ｂは、セグメントおよびその対応する冗長情報を記憶装置間でランダムに分散した一実施形態を示す。図１Ｂでは、ｗ，ｘ，ｙおよびｚで示す４つの記憶装置４２が、１，２，３および４で示す４つのセグメントに分割したデータを格納する。セグメントに対する冗長情報は、１つ以上のセグメントに基づくことができる。この例では、ここでは「冗長セット」と呼ぶものにおいて、２つのセグメントを用いる。冗長セットにおけるセグメントｉ，ｊの排他的ＯＲを計算して、冗長情報R_{ij}を得る。冗長情報R_{ij}およびセグメントｉの排他的ＯＲにより、セグメントｊを生成する。同様に、冗長情報R_{ij}およびセグメントｊの排他的ＯＲにより、セグメントｉを生成する。冗長セット内の各セグメントおよび冗長情報は、異なる記憶装置上に格納する。セグメントおよび冗長情報のランダム分散の一例を図１Ｂに示す。ここでは、セグメント３および４に対する冗長情報$R_{3,4}$を記憶装置ｗ上に格納し、セグメント２および３を記憶装置ｘ上に格納し、セグメント１を記憶装置ｙ上に格納し、セグメント４および冗長情報$R_{1,2}$を記憶装置ｚ上に格納する。また、フォールト・トレランスの技術分野においては公知である他の多くの技法を用いて、冗長情報を作成することも可能である。

【００３０】

冗長情報がセグメントのコピーである場合、セグメントのランダム分散は、セグメント・テーブル９０、または図２Ａに示すようなカタログに内において表わ

【００３２】

各セグメント・テーブルまたはファイル・マップは、他のセグメント・テーブルとは別個に格納することもできる。セグメント・テーブルは、カタログとして、一緒に格納することができる。カタログは、個々のクライアントにおいて、中央データベースにあるカタログ・マネージャ４９上に格納することができ、あるいはいくつかのデータベースまたはクライアント間で分散することも可能である。例えば、異なる種類のメディア・プログラム毎に、別個のカタログを維持することも可能である。例えば、放送ニュースの編成は、スポーツ・ニュース、天気予報、ヘッドライン・ニュース等に対して別個のカタログを有することができる。また、カタログは、他のデータと同様に記憶装置上に格納することも可能である。例えば、各クライアントは、乱数発生器のシードを用いて、カタログにアクセスすることができる。このようなカタログは、例えば、ネットワーク・ブロードキャスト・メッセージを全てのカタログ・マネージャまたはクライアントに送り、個々のセグメント・テーブルのカタログのコピーを得ることによって、他のクライアントが特定してデータにアクセスしたり、または復元要求を処理することができる。

【００３３】

データ・セグメントにアクセスするためには、各セグメントは一意の識別子を有していなければならない。セグメントのコピーは同じ一意の識別子を有することができる。２つ以上のセグメントに基づく冗長情報はそれ自体の識別子を有する。セグメントに対する一意の識別子は、ファイルのようなソースに対する一意の識別子、およびセグメント番号の組み合わせである。ソースまたはファイルに対する一意の識別子は、例えば、システム時間、またはデータをソースから捕獲した時に判定したその他の一意の識別子によって、あるいはファイルの作成時に決定することができる。以下で説明するが、ファイル・システムは、カタログ・マネージャにアクセスし、各ソース毎のセグメント・テーブル、またはセグメント識別子ならびにセグメントおよび冗長情報を格納してある記憶装置をリストに纏めたファイルを得ることができる。また、各記憶装置は、別個のファイル・システムも有することができ、これには、セグメント識別子、およびこれらを格納

してある記憶装置上の位置のディレクトリを収容する。クライアントが実行するアプリケーション・プログラムは、ソースまたはファイルの識別子、および恐らくは当該ソースまたはファイル内のある範囲のバイトを用いて、クライアントのファイル・システムからデータを要求することができる。すると、クライアントのファイル・システムは、一意のセグメント識別子を用いて、そのソースまたはファイルのセグメント・テーブルを突き止め、どのセグメントにアクセスする必要があるかについて判定を行い、各セグメント毎にデータを読み出す記憶装置を選択する。

【００３４】

再度図１Ａおよび図１Ｂを参照する。アプリケーション４４が記憶装置４２の１つにおいて選択したデータ・セグメントへのアクセスを要求する場合、記憶装置はその要求をキュー４８上に置く。キュー４８は、記憶装置が維持している。アプリケーションはこのような要求を、互いにまたはあらゆる中央制御部にも独立して行なうことができ、このためシステムは容易にスケーラブルとなっている。冗長情報がセグメントのコピーである場合、要求を送る先の記憶装置の選択は、多数の記憶装置４２上の多数のアプリケーション４４によって加えられる負荷のランダムな変動が、記憶装置４２全てにわたって統計的にかつ一層等しく均衡化されるように、制御することができる。例えば、アプリケーション４４からの各要求は、最も短い要求のキューを有する記憶装置によって処理することができる。あらゆる種類の冗長情報でも、アプリケーションおよび記憶装置間のデータの転送は、ネットワークの輻輳を低減するようにスケジュールすることができる。データ要求は、２段階で実行することができる。即ち、データをディスクから記憶装置上のバッファに転送する予備リード要求と、ネットワークを通じてバッファからアプリケーションにデータを転送するネットワーク転送要求である。これら２つの異なる要求を処理するために、キュー４８は、ディスク・キューおよびネットワーク・キューを含むとよい。

【００３５】

このランダムに分散したデータ・セグメントおよび対応する冗長情報の組み合わせ、ならびにネットワークを通じたデータ転送のスケジューリングにより、多

数の記憶装置および多数のアプリケーション間でスケーラブルにかつ信頼性高く双方向に、多数の独立した高帯域幅データ・ストリームを転送することができるシステムが得られる。セグメントのコピーを冗長データとして用いることにより、リード・アクセスのための記憶装置の選択は、記憶装置の相対的な負荷に基づいて行なうことができ、処理能力の向上を図ることができる。

【００３６】

次に図３を参照し、いくつかの記憶装置にランダムに分散してデータ・セグメントの多数のコピーを格納するプロセス例について、更に詳しく説明する。１つ以上のセグメントに基づく冗長情報を用いたプロセス例については、図２４に関連付けて以下で説明する。以下の説明は、モーション・ビデオ・データのリアル・タイム・キャプチャに基づくものである。この例は、オーディオのような他の時間的に連続するメディア、静止画像やテキストのような離散メディア、または知覚データのようなその他のデータをも含むがこれらには限定されない、他の形態のデータに一般化することができる。

【００３７】

米国特許第５，６４０６０１号および第５，５７７，１９０に記載されているように、どのようにしてリアル・タイム・モーション・ビデオ情報をコンピュータ・データ・ファイル取り込むかは、一般的に周知である。この手順は、取り込んだデータをセグメントに分割し、コピーし、セグメントのコピーを記憶装置間でランダムに分散するためのステップを含むように変更することも可能である。最初に、ステップ１２０において、キャプチャ・システムはセグメント・テーブル９０Ａ（図２Ａ）を作成する。典型的に、取り込むデータ・ストリームへのオフセットに、各画像をマップする画像インデックスも作成する。インデックス化したデータは、例えば、フィールドまたはフレームに対応することができる。インデックスは、オーディオのような別の種類のデータについては、時間期間のような他のサンプル境界を引用することができる。また、キャプチャ・システムは、利用可能な記憶装置のリストも取得する。どの記憶装置が利用可能かを識別する方法については、図１０ないし図１２に関連付けて以下で更に詳しく説明する。

【００３８】

ステップ１２１において、キャプチャ・システムはデータ・セグメントを作成する。セグメントのサイズは、例えば、モーション・ビデオ情報では１／４メガバイト、１／２メガバイト、または１メガバイトとするとよい。オーディオ情報は、例えば、１／４メガバイトのようなサイズを有するセグメントに分割するとよい。セグメント・サイズの格納および送信の分割に対する整合を得るためには、可能であれば、セグメントのサイズは、未圧縮または固定データ・レート、ディスク・ブロックおよびトラック・サイズ、メモリ・バッファ・サイズ、ネットワーク・パケット（例えば、６４Ｋ）および／またはセル・サイズ（例えば、ＡＴＭでは５３バイト）の倍数に関係付けるとよい。データが未圧縮かまたは固定レートの圧縮を用いて圧縮されている場合、セグメントは、時間的なサンプル境界で分割すれば、画像インデックスとセグメント・テーブルとの間に整合を得ることができる。概して言えば、セグメント・サイズは、システム・オーバーヘッドを低減するためには、大きくする方がよい。システム・オーバーヘッドは、セグメントが小さい程増大する。一方、記憶装置の全てにデータが分散されないような格納データ量およびセグメント・サイズである場合、コンボイ効果が発生する確率は高くなる。加えて、セグメント・サイズが大きい程、ディスク要求およびネットワーク要求を完了するためのレイテンシも増大する。

【００３９】

次に、ステップ１２２において、キャプチャ・システムは、選択したセグメントを格納するために利用可能な記憶装置のリストから、少なくとも２つの記憶装置４２を選択する。一方のセグメントのコピーのための記憶装置の選択は、ランダムまたは疑似ランダムである。この選択は、直前または直後のセグメントに対して行なう選択とは独立とすることも可能である。また、選択を行なった記憶装置の集合は、利用可能な記憶装置全ての部分集合とすることも可能である。１組の記憶装置の選択は、各ソースまたはファイル毎にランダムまたは疑似ランダムとすることもできる。この部分集合のサイズは、各記憶装置が少なくとも１つの異なるデータ・セグメントを有し、コンボイ効果の発生確度を極力低下させるようにしなければならない。即ち、データは、１組の記憶装置の数の少なくとも２

倍の長さ（セグメントにおいて）としなければならない。部分集合のサイズも、いずれの所与の時点においても、部分集合内の２つ以上の記憶装置が故障する、即ち、二重故障が発生する確度を低下させるためには、制限しなければならない。例えば、５つの内２つの記憶装置が故障し得る確率は、１００個の内２つの記憶装置が故障し得る確率よりも低く、したがって、データを分散する記憶装置の数を制限しなければならない。しかしながら、性能と部分集合のサイズとの間にはトレードオフがある。例えば、１００個の記憶装置から１０個の部分集合をランダムに選択して用いる場合、１００個の記憶装置の内２つが故障すると、ファイルの１０パーセントが悪影響を受ける。部分集合がないと、典型的に、ファイルの１００パーセントが悪影響を受けることになる。

【００４０】

二重故障、即ち、２つ以上の記憶装置が故障するという稀な確度では、データ・セグメントが失われる可能性がある。標準的なビデオ・ストリームでは、１セグメントの損失のために、プログラム素材１分において１または２フレームが失われる可能性がある。所与のソースまたはファイルに対するこのような故障の頻度は、その帯域幅および記憶装置数の関数である。即ち、

ｓ＝メガバイト（ＭＢ）単位での失われたデータのサイズ

ｎ＝記憶装置の初期数

ｂ＝ＭＢ単位での１秒当たりの記憶装置の平均帯域幅

ＭＴＢＦ＝故障間平均時間

ＭＴＴＲ＝故障または交換までの平均時間

ＭＴＤＦ＝二重故障不良の平均時間

ＳＭＴＢＦ＝故障間の全システム平均時間

とすると、

【００４１】

【数１】

$$SMTBF = \frac{MTBF}{n} \quad \text{および} \quad MTDF = \frac{1}{MTTR} * \frac{MTBF}{n} * \frac{MTBF}{(n-1)}$$

となる。

一例として、１００個の記憶装置を有し、各々５０ギガバイトの容量を有するシステムでは、ＭＴＴＲが１時間、ＭＴＢＦが１０００時間即ち６週間の場合、１１５年で二重故障不良が発生する確度となる。ＭＴＴＲが２４時間に増大すると、４．８年で二重故障不良が発生する確度となる。

【００４２】

再び図３を参照する。２つの記憶装置を選択した後、ステップ１２４において現セグメントを選択した記憶装置の各々に送り格納する。これらのライト要求は、順次ではなく、非同期とすることも可能である。次に、キャプチャ・システムは、ステップ１２６において、記憶装置がセグメントの格納完了を承認するのを待つ。捕獲しながらデータをリアル・タイムで格納する場合、ステップ１２４におけるデータ転送は、以下で更に詳しく論ずるリード動作と同様、２段階で行なうとよい。即ち、最初にクライアントが記憶装置に、データを格納するための空きバッファを準備するように要求することができる。記憶装置は、バッファが利用可能な推定時間を回答することができる。この推定時間に到達した場合、キャプチャ・システムは、記憶装置にデータを受け取るように要求することができる。次いで、記憶装置はそのバッファ内のデータを受け取り、そのバッファ内のデータをその記憶媒体に転送し、承認をキャプチャ・システムに送ればよい。

【００４３】

キャプチャ・システムが承認を受け取る前に、時間切れとなった場合、セグメントを再度同じ記憶装置または異なる記憶装置に送ることもできる。他のエラーも、キャプチャ・システムによって処理することができる。選択した記憶装置上での格納成功を確実にする動作を、セグメントの各コピー毎に別個のスレッドによって行なうことも可能である。

【００４４】

データを記憶装置に格納することに成功した後、ステップ１２７においてチャプチャ・システムはセグメント・テーブル９０を更新する。キャプチャが完了したとステップ１２８において判断した場合、プロセスは終了する。それ以外の場合、ステップ１２１に戻って、次のセグメントのためにプロセスを繰り返す。セ

グメント・テーブルは、例えば、キャプチャ・システムの主メモリ内に、ファイル・システムの一部として維持することができる。本例では、キャプチャ・システムはセグメント・テーブルおよび記憶装置の選択を管理するが、カタログ・マネージャ４９のようなシステムの他の部分が、これらのアクティビティを調整することも同様に可能である。ステップ１２９において、更新したセグメント・テーブルは、例えば、カタログ・マネージャに送ることができる。あるいは、カタログ・マネージャは、蓄積したシステム動作の知識を用いることによって、セグメント・テーブルを生成することもでき、更に要求に応じてこのテーブルをキャプチャ・システムに送ることもできる。

【００４５】

図４は、記憶装置がどのようにして捕獲したデータ・セグメントまたは冗長情報を格納するのかについて更に詳細に記載したフローチャートである。ステップ１４０において、記憶装置は、キャプチャ・システムからデータ・セグメントを受け取り、記憶装置のバッファにデータを格納する。記憶装置は格納のためにデータ・ファイルを用いると仮定すると、ステップ１４２において記憶装置はデータ・ファイルを開き、ステップ１４４においてデータ・ファイルにデータを格納する。カタログ・マネージャは、セグメントを格納するロケーションを指定することも可能である。データは、既存のデータ・ファイルに添付することができ、あるいは別個のデータ・ファイルに格納することもできる。先に論じたように、ステップ１４５において、記憶装置またはカタログ・マネージャは、各セグメント毎の一意の識別子を用いることにより、更にセグメント識別子を記憶装置上のそのロケーションにマップするテーブルを格納することにより、セグメントを追跡することも可能である。このテーブルは、記憶装置上におけるデータ・ファイルの抜き取り（ａｂｓｔｒａｃｔｉｏｎ）を実現することができる。記憶装置がいつデータを実際にその主ストレージに書き込むかは、他のアプリケーションで保留となっているその他のリードおよびライト要求に左右される。これら同時要求の管理については、以下で更に詳しく扱うことにする。次に、ステップ１４６において、ファイルを閉じることができる。ステップ１４８において、承認をキャプチャ・システムに送ることができる。

【００４６】

図３および図４のプロセスが完了したなら、捕獲したデータを、各セグメント毎に少なくとも２つのコピーを用いて、数個の記憶装置にランダムに分散する。多数のアプリケーションがこのデータへのアクセスを要求することができる。このアクセスを行なう態様は、恐らくランダムである。したがって、いずれの記憶装置も、多数のアプリケーションから、記憶装置上に格納してあるファイルからデータを読み出す要求およびファイルにデータを書き込む要求を多数受ける可能性があることは明白である。要求を管理するために、前述のように、要求キュー４８を記憶装置４２の各々によって維持する。実施形態の一例についての以下の説明では、記憶装置は２つのキュー、即ち、一方はディスク・アクセス要求のため、他方はネットワーク転送要求のために維持する。これらのディスクおよびネットワーク・キューの一実施形態については、以下で図１９に関連付けて更に詳しく説明する。

【００４７】

クライアント４４上で実行するアプリケーション・プログラムがデータを要求した場合、各データ・セグメントを少なくとも２つの記憶装置上に格納する際に、要求を満たすように記憶装置を選択する。要求データに対するセグメント・テーブル９０をこの目的のために用いる。記憶装置の選択は、データを要求するアプリケーション・プログラムによって、アプリケーション・プログラムを実行するクライアントのファイル・システムによって、記憶装置間の調整によって、またはカタログ・マネージャのような別のアプリケーションによって行なうことができる。選択はランダムでも疑似ランダムでもよく、あるいは最低使用頻度アルゴリズムに基づいて、または記憶装置のキューの相対的な長さに基づいて行なうことも可能である。利用可能な記憶装置上のキューの相対的長さに基づいて記憶装置を選択することによって、多数のアプリケーションの負荷を記憶装置の集合全体に一層等しく分散することが可能となる。このような選択については、図１６ないし図１８に関連付けて以下で更に詳しく説明する。

【００４８】

これより、特定的な一実施形態の詳細について更に説明する。この目的のため

には、記憶装置４２は、サーバまたは独立制御記憶装置として実現するとよく、アプリケーション４４をクライアントと呼ぶことにする。クライアントは種々のタスクを実行するアプリケーション・プログラムを実行することができる。サーバまたはクライアントを実現するのに適したコンピュータ・システムは、典型的に、主ユニットを含み、これは通常バスまたはスイッチのような相互接続機構を介してメモリ・システムに接続するプロセッサを含む。サーバおよびクライアント双方は、これらをコンピュータ・ネットワークに接続するネットワーク・インターフェースも有する。ネットワーク・インターフェースは、フォールト・トレランスに対応するために冗長化するとよい。また、クライアントは、ディスプレイのような出力デバイスや、キーボードのような入力デバイスを有することができる。入力デバイスおよび出力デバイスは双方とも、相互接続機構を介して、プロセッサおよびメモリ・システムに接続することができる。

【００４９】

尚、１つ以上の出力デバイスをクライアント・システムに接続してもよいことは理解されよう。出力デバイスの例には、陰極線管（ＣＲＴ）ディスプレイ、液晶ディスプレイ（ＬＣＤ）、プリンタ、モデムまたはネットワーク・インターフェースのような通信デバイス、ならびにビデオおよびオーディオ出力が含まれる。また、１つ以上の入力デバイスをクライアント・システムに接続してもよいことも理解されよう。入力デバイスの例には、キーボード、キーパッド、トラックボール、マウス、ペンおよびタブレット、モデムまたはネットワーク・インターフェースのような通信デバイス、ビデオおよびオーディオ・ディジタイザ、ならびにスキャナが含まれる。尚、本発明は、コンピュータ・システムと組み合わせて用いる特定の入力デバイスにも出力デバイスにも、更にはここに記載されるものにも限定されないことは理解されよう。

【００５０】

コンピュータ・システムは、「Ｃ」および「Ｃ＋＋」プログラミング言語のような、高級コンピュータ・プログラミング言語を用いてプログラム可能な汎用コンピュータ・システムとすることができる。また、コンピュータ・システムは、特別にプログラムした特殊目的ハードウエアとすることもできる。汎用コンピュ

ータ・システムでは、プロセッサは典型的に市販のプロセッサであり、その例には、Ｉｎｔｅｌ（インテル社）から入手可能な、ＭＭＸ技術を用いたＰｅｎｔｉｕｍ ＩＩのようなｘ８６シリーズ・プロセッサや、ＡＭＤやＣｙｒｉｘから入手可能な同様のデバイス、Ｍｏｔｏｒｏｌａから入手可能な６８０Ｘ０シリーズ・マイクロプロセッサ、Ｄｉｇｉｔａｌ Ｅｑｕｉｐｍｅｎｔ Ｃｏｒｐｏｒａｔｉｏｎ（ディジタル・エクイップメント社）から入手可能なＡｌｐｈａシリーズ、およびＩＢＭから入手可能なＰｏｗｅｒＰＣプロセッサがある。多くのその他のプロセッサも利用可能である。このようなマイクロプロセッサは、オペレーティング・システムと呼ばれるプログラムを実行することができ、その例には、ＷｉｎｄｏｗｓＮＴ、Ｗｉｎｄｏｗｓ９５、ＵＮＩＸ、ＩＲＩＸ、Ｓｏｌａｒｉｓ、ＤＯＳ、ＶＭＳ、ＶｘＷｏｒｋｓ、ＯＳ／Ｗａｒｐ、ＭａｃＯＳ Ｓｙｓｔｅｍ ７およびＯＳ８オペレーティング・システムがある。オペレーティング・システムは、他のコンピュータ・プログラムの実行を制御し、スケジューリング、デバッギング、入出力制御、コンパイル、ストレージ割り当て、データ管理およびメモリ管理、ならびに通信制御および関連サービスを行なう。プロセッサおよびオペレーティング・システムは、高級プログラミング言語でアプリケーション・プログラムを各ためのコンピュータ・プラットフォームを定義する。

【００５１】

各サーバは、大量の主メモリ、例えば、３２メガバイトよりも遥かに多いメモリや、例えば、数ギガバイトのディスク容量を有する安価なコンピュータを用いて実現することができる。ディスクは、１つ以上の単体ディスク、または独立したディスクの冗長アレイ（ＲＡＩＤ）、あるいはその組み合わせとすることができる。例えば、サーバは、ＷｉｎｄｏｗｓＮＴまたはＶｘＷｏｒｋｓのようなリアル・タイム・オペレーティング・システムを有する、Ｐｅｎｔｉｕｍまたは４８６マイクロプロセッサを用いたシステムとすることができる。オーサリング・システム、キャプチャ・システム、および再生システムは、この主の製品のために当技術分野において現在用いられているプラットフォームを用いて実現することができる。例えば、マサチューセッツ州ＴｅｗｋｓｂｕｒｙのＡｖｉｄ Ｔｅｃｈｎｏｌｏｇｙ，Ｉｎｃ．（アビッド・テクノロジ社）からのＭＥＤＩＡＣＯ

ＭＰＯＳＥＲは、ＰｏｗｅｒＰＣマイクロプロセッサおよびＭａｃＯＳ Ｓｙｓｔｅｍ ７オペレーティング・システムを有するＡｐｐｌｅ Ｃｏｍｐｕｔｅｒ，Ｉｎｃ．（アップル・コンピュータ社）からのＰｏｗｅｒ Ｍａｃｉｎｔｏｓｈコンピュータを用いている。ＷｉｎｄｏｗｓＮＴオペレーティング・システムを備え、ＩｎｔｅｌからのＭＭＸ技術を用いたＰｅｎｔｉｕｍＩＩプロセッサに基づくシステムも使用可能である。再生システムの例には、コロラド州ＢｏｕｌｄｅｒのＰｌｕｔｏ Ｔｅｃｈｎｏｌｏｇｉｅｓ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｉｎｃ．（プルート・テクノロジーズ・インターナショナル社）からの「ＳＰＡＣＥ」システム、またはＭａｃｈｉｎｔｏｓｈプラットフォームを用いたＡｖｉｄ ＴｅｃｈｎｏｌｏｇｙからのＡＩＲＰＬＡＹシステムが含まれる。カタログ・マネージャは、Ｉｎｆｏｒｍｉｘデータベースのような、適当なデータベース・システムに対応するいずれかのプラットフォームを用いて実現することができる。同様に、システム内で利用可能なデータの種類を追跡するアセット・マネージャも、このようなデータベースを用いて実現することができる。

【００５２】

コンピュータにおけるメモリ・システムは、典型的に、コンピュータ読み取り可能および書き込み可能不揮発性記録媒体を含み、その例には、磁気ディスク、光ディスク、フラッシュ・メモリおよびテープがある。ディスクは、フロップ・ディスクまたはＣＤ－ＲＯＭのようなリムーバブルでも、ハード・ドライブのように固定でもよい。ディスクは多数のトラックを有し、その中に、典型的に二進形態、即ち、一連の１および０として解釈した形態で、信号を格納する。このような信号は、マイクロプロセッサが実行するアプリケーション・プログラム、またはディスク上に格納してありアプリケーション・プログラムが処理する情報を定義することができる。典型的に、動作において、プロセッサは不揮発性記録媒体から集積回路メモリ素子にデータを読み出す。集積回路メモリ素子は、典型的に、ダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ）またはスタティック・メモリ（ＳＲＡＭ）のような、揮発性ランダム・アクセス・メモリである。集積回路メモリ素子は、ディスクよりも、プロセッサによる情報への速いアクセスを可能とする。プロセッサは通常集積回路メモリ内でデータを操作し、次いで

処理が完了すると、データをディスクにコピーする。ディスクと集積回路メモリ素子との間のデータ移動を管理する機構には、様々なものが公知であり、本発明はそれに限定されるものではない。また、本発明は特定のメモリ・システムにも限定される訳ではないことも理解されよう。

【００５３】

本発明は、特定のコンピュータ・プラットフォーム、特定のプロセッサ、特定の高級プログラム言語にも限定される訳ではないことは理解されよう。加えて、コンピュータ・システムは、マイクロプロセッサ・コンピュータ・システムとしてもよく、あるいはコンピュータ・ネットワークを通じて接続した多数のコンピュータを含むことも可能である。

【００５４】

前述のように、各記憶装置４２は、サーバを通じてアクセスされた場合、各アプリケーション４４はファイル・システムを有することができる。ファイル・システムは、典型的に、オペレーティング・システムの一部であり、データのファイルを維持する。ファイルは、有名論理構造であり、ファイル・システムによって定義されかつ実現され、データの論理レコードの名称およびシーケンスを、物理的なストレージ・メディア上のロケーションにマップする。ファイル・システムは、データの物理的ロケーションをアプリケーション・プログラムからマスクするが、ファイル・システムは通常隣接するブロック内にある１つのファイルのデータを物理的ストレージ・メディアに格納しようとする。ファイルは、種々のレコード型に特定して対応することができ、あるいは未定義として、アプリケーション・プログラムに解釈または制御させることも可能である。ファイルは、その名称またはその他の識別子でアプリケーション・プログラムによって呼ばれ、オペレーティング・システムが定義するコマンドを用いて、ファイル・システムを通じてアクセスされる。オペレーティング・システムは、ファイルを作成し、ファイルを開き、ファイルに書き込み、ファイルを読み取り、ファイルを閉じるための基本的ファイル動作を行なう。これらの動作は、ファイル・システムに応じて、同期でも非同期でもよい。

【００５５】

ここで記載する場合、ファイルまたはソースのデータはセグメント単位で格納し、そのコピーまたはその他の形態の冗長情報を多数の記憶装置の間でランダムに分散する。

【００５６】

殆どのファイル・システムについて概して言えば、ファイルを作成するためには、オペレーティング・システムは最初に、ファイル・システムが制御するストレージ内において、空間を特定する。次に、新たなファイルに対するエントリをカタログ内に作成する。カタログは、利用可能なファイルの名称およびファイル・システム内におけるそれらのロケーションを示すエントリを含む。ファイルの作成は、一定の利用可能な空間をファイルに割り当てることを含むこともある。一実施形態では、ファイルに対してセグメント・テーブルを作成することができる。ファイルを開くと、典型的に、ハンドルがアプリケーション・プログラムに戻され、これを用いてそのファイルにアクセスする。ファイルを閉じると、ハンドルは無効となる。ファイル・システムはハンドルを用いて、ファイルに対するセグメント・テーブルを識別する。

【００５７】

ファイルにデータを書き込むために、アプリケーション・プログラムはコマンドをオペレーティング・システムに発行し、オペレーティング・システムは、ファイル名、ハンドルまたはその他の記述子のようなファイルのインディケータ、およびファイルに書き込む情報の双方を指定する。概して言えば、ファイルのインディケータが与えられると、オペレーティング・システムはディレクトリを探索して、ファイルのロケーションを見つけ出す。データは、ファイル内の既知のロケーションまたはファイルの終端に書き込むことができる。ディレクトリ・エントリは、ライト・ポインタと呼ばれる、ポインタを現在のファイル終端に格納することができる。このポインタを用いて、ストレージの次に利用可能なブロックの物理的ロケーションを計算し、そのブロックに情報を書き込むことができる。ライト・ポインタをディレクトリ内において更新し、新たなファイル終端を示すことができる。一実施形態では、ライト動作は、ファイルのセグメントのコピーを記憶装置間でランダムに分散し、ファイルに対するセグメント・テーブルを更

新する。また、ライト動作も、セグメントおよび対応する冗長情報を異なる記憶装置に格納することができる。

【００５８】

ファイルからデータを読み出すためには、アプリケーション・プログラムは、オペレーティング・システムにコマンドを発行し、ファイルのインディケータ、およびアプリケーションに割り当てられているメモリ・ロケーションの内リード・データが位置するメモリ・ロケーションを指定する。概して言えば、ファイルのインディケータが与えられると、オペレーティング・システムはそのディレクトリを探索して関連するエントリを求める。アプリケーション・プログラムは、用いるファイルの開始からのオフセットを指定することができ、あるいはシーケンシャル・ファイル・システムでは、ディレクトリは次に読み取るデータ・ブロックへのポインタを与えることができる。一実施形態では、記憶装置の選択やデータ転送のスケジューリングは、クライアントのファイル・システムのリード動作の一部として実現する。

【００５９】

クライアントは、既定のアプリケーション・プログラミング・インターフェース（ＡＰＩ）と共にファイル・システムまたは特殊コード・ライブラリを用いて、ファイルの一部に対する要求を、選択した記憶装置からのデータ・セグメントに対する要求に変換する。記憶装置は、クライアントのファイル・システムとは完全に分離した、それ自体のファイル・システムを有することも可能である。記憶装置上のセグメントは全て、例えば、記憶装置の単一ファイル内に格納することができる。あるいは、クライアント・ファイル・システムは、ネットワーク上で記憶装置を生ストレージ（ｒａｗ　ｓｔｒａｇｅ）として用い、カタログ・マネージャおよびセグメント・テーブルを用いてファイルの抜き取りを実現することも可能である。ファイルのセグメント・テーブルは、各セグメントに選択した記憶装置上における当該セグメントのロケーションも示すことができる。

【００６０】

ファイル・システムを用いる主要な利点の１つは、アプリケーション・プログラムにとっては、ファイルは、物理的な記憶媒体やオペレーティング・システム

がデータを格納するために用いる当該媒体上のロケーションについて何の懸念もなく、作成し、開き、書き込み、読み出し、そして閉じることができる論理構造であるという点にある。ネットワーク・ファイル・システムでは、ファイル・システムは、種々の記憶装置からの指定ファイルからのデータ要求を管理し、アプリケーション・プログラムは、データが格納されている物理的ストレージまたはコンピュータ・ネットワークに関する詳細を知る必要は全くない。記憶装置がそれ自体の独立したファイル・システムを有する場合、クライアント・ファイル・システムも、記憶装置の記憶機構の詳細を知る必要はない。記憶装置は、例えば、ＷｉｎｄｏｗｓＮＴファイル・システムと関連するファイル・システム、またはＶｘＷｏｒｋｓのようなリアル・タイム・オペレーティング・システムのファイル・システム、または非同期動作を許すファイル・システムを用いることができる。

【００６１】

記憶装置は、クライアント、およびオプションとしてコンピュータ・ネットワークを用いてカタログ・マネージャと相互接続されている。コンピュータ・ネットワークは、１組の通信チャネルであり、互いに通信可能な１組のコンピュータ・デバイスまたはノードを相互接続する。ノードは、クライアントのようなコンピュータ、記憶装置およびカタログ・マネージャ、またはスイッチ、ルータ、ゲートウェイおよびその他のネットワーク・デバイスのような様々な種類の通信デバイスとすることができる。通信チャネルは、光ファイバ、同軸ケーブル、銅の撚り線対、衛星リンク、ディジタル・マイクロ波無線等を含む種々の伝送媒体を用いることができる。

【００６２】

コンピュータ・ネットワークは、ネットワークによるノードの接続の幾何学的構成であるトポロジを有する。トポロジの種類には、二点間接続、線形バス、環状接続、星形結線、および多重接続ネットワークが含まれる。ネットワークは、これら基本的なトポロジの様々な組み合わせを用いることができる。トポロジは、物理的な設置に応じて変化する可能性がある。各ノード、即ち、クライアントまたは記憶装置を直接同じスイッチに接続する、遮断のないスイッチに基づくネ

ットワークを用いることができる。実施態様によっては、多数のクライアントおよび記憶装置を物理的ループまたはサブネットワークに接続し、これらを切り替え構造（ｓｗｉｔｃｈｉｎｇ　ｆａｂｒｉｃ）に相互接続することができる。また、本システムは、多数のスイッチを用いて接続することも可能である。

【００６３】

また、ネットワークは、プロトコル、メッセージ・フォーマット、および通信ハードウエアおよびソフトウエアがネットワーク上のデバイス間で通信を行なうためのその他の規格を規定するネットワーク・アーキテクチャも有する。一般的に用いられるネットワーク・アーキテクチャは、Ｏｐｅｎ　Ｓｙｓｔｅｍｓ　Ｉｎｔｅｒｃｏｎｎｅｃｔｉｏｎ　ｒｅｆｅｒｅｎｃｅ　ｍｏｄｅｌ（オープン・システムズ相互接続基準モデル）として知られている、Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｓｔａｎｄａｒｄｓ　Ｏｒｇａｎｉｚａｔｉｏｎ　ｓｅｖｅｎ－ｌａｙｅｒ　ｍｏｄｅｌ（国際標準機構７レイヤ・モデル）である。７レイヤとは、アプリケーション、プレゼンテーション、セッション、トランスポート、ネットワーク、リンクおよび物理レイヤである。各機械は、これらの層の１つにおいて、同じ通信プロトコルを用いて他のあらゆる機械と通信を行なう。

【００６４】

一実施形態では、リンク層は、好ましくは、パッケージがクライアントに受信される順序を保持し、無限のレイテンシの潜在的可能性を回避するようにしたものである。したがって、適当なリンク・レイヤ・プロトコルはＯＣ３、ＯＣ１２、または更に高い帯域幅ネットワークのような、非同期転送モード（ＡＴＭ）ネットワークを含む。ＡＡＬ１５モードで動作するＡＴＭシステムが好ましい。１００Ｔｘないしギガビット（１，０００Ｔｘ）の容量を有するイーサネット・ネットワークも、ソースから宛先に効率的なパケット伝送を提供する。適当なイーサネット・ネットワーク・プラットフォームは、例えば、カリフォルニア州、Ｓａｎｔａ　Ｃｌａｒａの３Ｃｏｍから入手可能である。ＡＴＭシステムの一例は、ペンシルベニア州、ＷａｒｒｅｎｄａｌｅのＦｏｒｅ　Ｓｙｓｔｅｍｓ（フォア・システムズ社）、またはマサチューセッツ州、ＣｏｎｃｏｒｄのＧｉｇａ－Ｎｅｔ（ギガネット社）から入手可能である。ＦｉｂｒｅＣｈａｎｎｅｌ、ＦＤ

ＤＩ、またはＨＩＰＰＩネットワークも使用可能である。異なるクライアント、カタログ・マネージャおよび記憶装置は全て、リンク・レイヤ・プロトコルを用いて通信することができる。また、このレイヤにおける通信は、各レイヤのプロトコル毎にカプセル化したデータを処理するために実行するメモリ・コピーによるオーバーヘッドを低減する。マサチューセッツ州、ＴｙｎｇｓｂｏｒｏのＰｏｌｙｂｕｓ　Ｓｙｓｔｅｍｓ　Ｃｏｒｐｏｒａｔｉｏｎ（ポリバス・システムズ社）からの帯域幅分散ネットワーク・ファイル・システムも使用可能である。

【００６５】

これまで一実施形態のコンピュータ・プラットフォームについて説明してきたが、以下では一実施形態の動作および詳細について更に追加説明を行なう。

【００６６】

一実施形態では、記憶装置および当該記憶装置上に格納してあるデータを維持するためのプロセスがある。例えば、故障回復プロシージャは、ファイルの追加コピーの作成を含む場合がある。加えて、ファイルの可用性、即ち、ファイルに対するアクセスの信頼性の必要性に基づいて、ファイルを削除または追加することも可能である。最後に、保守手順によっては、記憶装置上のファイル削除、別の記憶装置へのファイル・コピー、およびシステムからの記憶装置の除去を伴う場合もある。また、ファイルは、保管したり、あるいはシステムからアーカイブ・システムに移動させることも可能である。これらのプロセスについて、図５ない図９に関連付けて更に詳しく説明する。このようなデータ管理プロセスは、カタログ・マネージャ、別の記憶装置、またはクライアントによって行なうことができる。これらのプロセスをクライアントが行なう場合でも、カタログ・マネージャまたは記憶装置のリソースを占有することはなく、クライアントのデータ要求に対して回答するというような、他の重要性が高いタスクに用いることができる。

【００６７】

図５は、記憶装置が利用不能となり、その故障を検出した後にどのようにして故障回復を行なうことができるかについて、更に詳細に記載したフローチャートである。このような故障を検出する１つの方法について、図１０ないし図１２に

関連付けて以下で更に詳しく説明する。要求に対して繰り返し応答し損なうことを、故障を示すために用いることもできる。このプロセスの成功は、システムないにある各セグメントのコピー数、または冗長集合内のセグメント数に依存する。コピー数をＮとすると、Ｎ－１個の記憶装置が故障しても、システムはデータ損失なく動作する。ある記憶装置が故障した後、新たな記憶装置を代わりに設置し、失われたデータを再現するか、または失われたデータを再度作成し、残りの記憶装置全てに分配することができる。図５は、冗長データがセグメントのコピーである場合のプロセスを記載する。以下で説明する図２５は、冗長情報が２つ以上のセグメントに基づく場合のプロセスを示す。

【００６８】

ステップ２００において、データの追加コピーを、最初にデータ、例えば、再現すべきファイルまたはソースを選択することによって行なうことができる。再現するファイルは、優先順序によって選択することができ、更に自動的または手作業のいずれかで選択することも可能である。この種の復元は、いくつかのファイルからのデータを再構成し、他のファイルからのデータが復元される前に利用可能とすることができる。ステップ２０２において、ソースのセグメント・テーブルを用いて、失われたデータ・セグメント、即ち、失われた記憶装置上に格納してあったデータを特定する。ステップ２０４において、典型的にデータを最初に捕獲するときと同様に、故障した記憶装置と交換するために新たな記憶装置が利用できない場合、失われたセグメント毎に新たな記憶装置を選択する。あるいは、交換記憶装置を選択する。ステップ２０６において、失われたセグメントのコピーを代替記憶装置から読み出し、選択した記憶装置に格納する。ステップ２０４ないし２０８のファイル動作は、非同期とし、各セグメント毎に別個のスレッドによって実行することができる。このような動作は、このネットワーク・アーキテクチャに備えた、多数対多数のリード／ライト機能を利用する。次に、ステップ２０８において、コピー動作の完了成功次に、ファイルのセグメント・テーブルを更新する。プロセスが完了した場合、カタログ・マネージャがセグメント・テーブルを維持するのであれば、ステップ２０９において、新たなセグメント・テーブルを用いてカタログ・マネージャを更新することができる。基のセグ

メント・テーブルが、疑似ランダム・シーケンス発生器へのシードによって表わされる場合、実際のテーブルを作成し変更しなければならない場合がある。

【００６９】

このプロセスを用いた未ロード・システムに対する再入力（ｒｅｐｏｐｕｌａｔｉｏｎ）の速度および冗長性復元は、次の式で定義する。

【００７０】

【数２】

$$\frac{s}{(n-1+d)(b/2)}$$

ここで、ｓ＝メガバイト単位（ＭＢ）で示す、失われたデータのサイズ

ｎ＝記憶装置の初期数

ｂ＝ＭＢ／秒で表わす、記憶装置の平均帯域幅

ｄ＝ＭＢ／秒で表わす、ユーザ・デマンドの負荷

例えば、１０個の記憶装置の内１つが故障したために５０ギガバイトのストレージへのアクセスが失われた場合、ｎ＝１０記憶装置、単位帯域幅ｂ＝１０ＭＢ／秒とすると、（ｎ－１）＝９および（ｂ／２）＝５となる。したがって、他に負荷がなければ、復元には約２０分を要する。この絶対復元速度は、通常、クライアントへの変動再生負荷の逆数として減少する。例えば、５０％負荷の場合、入力時間は２００％増大する。呼び出されると、再分散タスクは速いレートで実行することができ、多元記憶装置チェッカボードを多数の記憶装置に切り替えるが、再入力アクティビティは適宜クライアント・ファイル・サービス要求にしたがって行われる。実際の影響は、故障した記憶装置によって記憶装置の全帯域幅に多少の損失が出る程度である。復元のためのファイル選択の優先順位を決定することにより、最も重要なファイルを真っ先に復元することを保証する。

【００７１】

図６は、どのようにしてデータの追加コピーを作成することができるかについて更に詳細に記載したフローチャートである。このプロセスを呼び出すと、ミッション・クリティカル（ｍｉｓｓｉｏｎ　ｃｒｉｔｉｃａｌ）な即ち要求の多い

データの追加データ・コピーが得られるようになる。新たなコピーに日付スタンプを与え、いつそのコピーを削除できるかを示すこともできる。選択したデータが与えられると、ステップ２１０において、データのセグメントを選択する。ステップ２１２において、各セグメントには、新たな記憶装置がランダムに割り当てられ、各記憶装置が所与のセグメントのコピーを多くても１つ有することを保証する。次に、ステップ２１４において、選択した記憶装置上にセグメントを格納する。このセグメントの格納完了に成功した場合、ステップ２１６においてデータのセグメント・テーブルを更新する。データのセグメント全てを未だコピーしていないとステップ２１７で判定した場合、プロセスを繰り返し、ステップ２１０に戻って、次のデータ・セグメントを選択する。プロセスが完了した場合、カタログ・マネージャがセグメント・テーブルを有しているのであれば、ステップ２１８において、新たなセグメント・テーブルによってカタログ・マネージャを更新することができる。このプロセスはセグメント全体にわたって連続的であるが、各セグメントは、別個のスレッドを用いて処理することができ、ステップ２１４のファイル動作は非同期とすることもできる。このような処理によって、コピーを素早く作成することを保証する。この手順により、疑似乱数発生器のシードを用いて、なおもセグメント・テーブルを表わすことができる。

【００７２】

図７は、データのコピーをどのようにして削除するのかについて詳細に記載したフローチャートである。このプロセスは、例えば、データがもはやさほど要求されなくなった場合に呼び出すことができる。例えば、コピー上の日付スタンプを用いて、いつデータを削除するかを示すことができる。所与のデータに対して図２に示したセグメント・テーブルを仮定すると、ステップ２２０において、コピー集合の１つ、即ち、テーブル内の１列を選択する。ステップ２２２において、列内の各セグメントを削除する。ステップ２２２において各セグメント毎に削除動作完了に成功した場合、ステップ２２４においてセグメント・テーブルを更新する。ステップ２２２および２２４は、セグメント毎に繰り返す。このプロセスは、セグメント全体を通じて連続的でもよく、または各セグメントを別個のスレッドによって処理することも可能である。プロセスが完了した場合、カタログ

・マネージャがセグメント・テーブルを維持しているのであれば、ステップ２２６において、カタログ・マネージャを新たなセグメント・テーブルを用いて更新することができる。

【００７３】

図８は、別のアクティブな記憶装置をどのようにしてシステムから除去するのかについて記載したフローチャートである。例えば、そのファイル・システムを用いて、そのファイルのリストを識別することによって、記憶装置上で利用可能なデータを識別する。最初に、記憶装置を、新たなセグメントを書き込むために利用可能とする。このステップは、例えば、カタログ・マネージャに通知することによって、または同報通信メッセージを全クライアントに送ることによって実行することができる。各ファイルのセグメントを他の記憶装置上に再分散し、その後に記憶装置をシステムから除去する。このファイル・リストが与えられると、ステップ２３０において、処理する次のファイルを選択する。ステップ２３２において、セグメント・テーブルを用いて、冗長情報を収容するセグメントを含む、記憶装置上のこのファイルのセグメント全てを識別する。ステップ２３４において、次に処理するセグメントを選択する。ステップ２３５において、残りの記憶装置からのランダム選択により、選択したセグメントに新たな記憶装置を割り当て、所与のセグメントのコピーを１つよりも多く有する記憶装置がないことを保証する。次に、ステップ２３６において、新たに選択した記憶装置にデータを書き込む。このライト動作の完了に成功した場合、ステップ２３７において、セグメント・テーブルを更新する。所与のファイルのセグメント全てを再分散したとステップ２３８において判定した場合、ステップ２３９において、セグメント・テーブルを適宜カタログ・マネージャに送ることができる。セグメントは、順次または非同期ファイル処理を用いる別個のスレッドによって処理することができる。カタログ・マネージャを更新した後に、古い記憶装置からセグメントを削除することができる。ステップ２４０において次のファイルがあると判定されたなら、処理は次のファイルに進む。全てのファイルを再分散し終えた場合、このプロセスは完了し、記憶装置をシステムから除去することができる。

【００７４】

図９は、保管またはバックアップのためにデータをどのようにコピーすることができるかについて記載するフローチャートである。このプロセスは、利用可能な記憶装置からのデータの各セグメントの１つのコピーを、アーカイブ記憶媒体のようなバックアップ記憶システムにコピーすることを伴う。各コピー・セットおよびあらゆる冗長情報も、全ての記憶ユニットから削除することができる。このプロセスは、ステップ２５０において、コピー・セット、例えば、Ａリストをセグメント・テーブルのコラムから選択することによって実行することができる。あるいは、前述のように他のアプリケーションによって適用される技法を用いて、各セグメントを順に読み出し、各セグメント毎に記憶装置の選択を行なうことも可能である。ステップ２５２において、選択したコピー・セットからの各セグメントは、その記憶装置から読み出し、記憶媒体上に格納する。各セグメントを記憶媒体にコピーするのに成功したなら、ステップ２５４において、全ての残りのコピー・セットまたはあらゆる冗長情報からの残りのセグメント全てを、記憶装置から削除することができる。セグメントは、順次処理することも、非同期ファイル処理を用いて別個のスレッドによって処理することも可能である。次に、ステップ２５６において、カタログ・マネージャを更新することができる。

【００７５】

図１０ないし図１２に関連付けて、記憶装置を監視してどのようにして可用性を判断し、故障を検出するかについてこれより説明する。記憶装置が利用可能か否かについて判定するには、いくつかの方法があり、記憶装置をポールすること、記憶装置からの例外を処理すること、または記憶装置が周期的にアプリケーションまたは複数のアプリケーションにその可用性について通知することを含む。一実施形態では、カタログ・マネージャ４９またはその他のいずれかのクライアントに加えて、双方がシステム内においてどの記憶装置４２がアクティブであるか監視し、各ファイル毎にセグメント・テーブルのカタログを維持することができる。記憶装置を監視する方法の１つを図１０ないし図１２に示す。システム上で利用可能な各記憶装置は、カタログ・マネージャにそれが利用可能であることを周期的に知らせるプロセスを確立する。即ち、このプロセスは、記憶装置が周期的にシステム・タイマからのタイマ割込またはイベントに応答してカウンタを

増分する、第１段階６０を有する状態機械と見なすことができる。このカウンタが、１００ミリ秒というような、ある所定量に到達した場合、別の状態６２への遷移が発生する。状態６２への遷移において、「ピング」と呼ぶ信号を記憶装置がカタログ・マネージャに送る。この信号は、１ＡＴＭセル程の小さいなメッセージとし、送信するために多くの帯域幅を用いないことが好ましい。この信号は、記憶装置の識別子、ならびに可能性として容量、効率および／または記憶装置の帯域幅可用性というような他の情報も含むことができる。次のタイマ割込またはイベント時に、カウンタをリセットし、状態６０に戻る遷移が行われる。

【００７６】

カタログ・マネージャは、利用可能な記憶装置を追跡するとよい。この目的のために、カタログ・マネージャは記憶装置のリスト７０を用いることができる。その一例を図１１に示す。この記憶装置リストは、７２に示す記憶装置に識別子でインデックスしたテーブルとして実施することができる。記憶装置が存在する即ち利用可能な場合、帯域幅、メモリ容量、または記憶装置のパワーに関するその他の情報を、列７４において得られるようにする。列７６に示すように、記憶装置からの最後の「ピング」以来のカウントもある。このカウントが、３００ミリ秒というような所定量を超過した場合、記憶装置は使用不能と見なされ、続いて前述のような故障回復手順を行なうことができる。記憶装置のリスト７０を維持する追跡プロセスの一例について、図１２に関連付けてこれより詳細に説明する。

【００７７】

図１２は、カタログ・マネージャがどの記憶装置が利用可能化について判定を行なう際に実行することができる追跡プロセスを記載した状態機械である。これらの状態機械の１つは、カタログ・マネージャ上のプロセスとして、各記憶装置毎に確立することも可能である。第１状態８０は、待ち状態であり、記憶装置のリスト７０内の記憶装置に対するカウント値７６を、周期的なタイマ割込に応答して、当該記憶装置に対して増分する。記憶装置から「ピング」を受信した場合、状態８２への遷移が起こる。状態８２において、リスト７０内のこの記憶装置の存在を検証する。記憶装置がリスト７０内にある場合、この記憶装置のカウン

ト７６をリセットし、この記憶装置に関する情報を更新することができ、そして状態８０に戻る遷移が行われる。記憶装置がリストにない場合、これをリストに追加し、カウントをリセットし、状態８０に戻る遷移が行われる。所与の増分の後、記憶装置のカウントが、３００ミリ秒というような所定のタイム・アウト値よりも大きくなった場合、故障回復手順を実行する。即ち、リスト７０から記憶装置を除去し、状態８４において、フォールト・トレラント手順を実行する。記憶装置からの「ピング」をカタログ・マネージャが受信し、その記憶装置が対応する追跡プロセスを有していない場合、カタログ・マネージャは記憶装置をリストに追加し、この記憶装置に対して追跡プロセスを作成する。

【００７８】

カタログ・マネージャ４９を有することに加えて、システムは、アセット・マネージャと呼ぶデータベースも含むことができる。アセット・マネージャは、各ファイル毎のインデックスのように、システムにおいて利用可能なメディア・ソースに関する種々の情報を格納する。カタログ・マネージャおよびアセット・マネージャは組み合わせることも可能である。アセット・マネージャに格納する有用な一種の情報は、図１３に示すテーブルであり、これは、ソース識別子および、米国特許第５，２６７，３５１号に示されるような、当該ソース内における範囲に基づいて、等価データ・ファイルを関係付けるものである。ソース識別子は、データの原ソースの指示であり、アナログ・ソースとすることができる。これに対して実際に利用可能なデータは、記憶装置上に格納してある当該ソースのディジタル化したコピーである。即ち、テーブルは、ソース識別子のエントリ、ソース識別子１０２内の範囲、および、データ・ファイルのリストのような、当該ソースからの等価なデータの指示１０４を有する。リスト１０４は、ソースに対するデータ・ファイルの１つを特定し次いで当該ファイルに対するセグメント・テーブルにアクセスし、データのセグメントが種々の記憶装置上のどこに分散されたかについて判定するために用いることができる。図２Ａのセグメント・テーブル９０Ａは、１０６および１０８に示すように、図１３のこのリストに組み込むこともできる。図２Ｂのセグメント・テーブル９０Ｂも同様に、リスト１０４に組み込むことができる。このようなデータ間の等価性は、あらゆるアプリケー

ション・プログラムによっても維持することができる。

【００７９】

カタログ・マネージャは、種々の記憶装置上にどのようにデータを分散したかを監視するデータベースであるので、フォールト・トレランスおよび可用性を高め、ボトルネックのなるその確度を低下するように設計して当然である。したがって、カタログ・マネージャは、従来の分散データベース管理技法を用いて実現する。また、Ｍｒａｔｈｏｎ　Ｔｅｃｈｏｎｏｌｏｇｉｅｓ（マラソン・テクノロジ社）、Ｔａｎｄｅｍ　Ｃｏｍｐｕｔｅｒｓ（タンデム・コンピュータズ社）、Ｓｔｒａｔｕｓ（ストラタス）、およびＴｅｘａｓ　Ｍｉｃｒｏ，Ｉｎｃ．（テキサス・マイクロ社）からのもののような、非常に可用性が高い機械を用いて、カタログ・マネージャを実現することも可能である。また、別個のクライアント・アプリケーションが用いるカタログ・マネージャもいくつかあり得る。あるいは、各クライアント・アプリケーションは、標準的な技法を用いて多数のデータ・コピー間の一貫性を維持し、それ自体のカタログのコピーをローカルに維持することも可能である。このように、カタログ・マネージャは、故障の中心点ではない。また、クライアントは、それ自体のカタログ・マネージャとして作用することも可能である。また、カタログは、データとして扱うことも可能であり、そのセグメントおよび冗長データを記憶装置間でランダムに分散する。各クライアントは、各カタログ毎に、セグメント・テーブル、または、セグメント・テーブルを表わす乱数発生器のシードを有することもできる。

【００８０】

以上、どのようにしてデータを捕獲し記憶装置に格納することができるか、そしてどのようにして記憶装置上のデータ格納を管理することができるかについて説明したので、これよりオーサリングおよび再生を行なうクライアント・アプリケーションについて、図４および図１５に関連付けて更に詳しく説明する。

【００８１】

マルチメディア・データのオーサリング、処理および表示を行なうには、いくつかの種類のシステムを用いることができる。これらのシステムは、データを修正し、異なるデータの組み合わせを定義し、新たなデータを作成し、データをユ

ーザに表示する際に用いることができる。これらの種類のシステムを実現する種々の技法が、当技術分野では公知である。

【００８２】

マルチメディア・オーサリング、処理および再生システムは、典型的に、マルチメディア組成（ｍａｒｕｔｉｍｅｄｉａ　ｃｏｍｐｏｓｉｔｉｏｎ）を表わすデータ構造を有する。このデータ構造は、最終的に、一意の識別子またはファイル名のようなソース素材の識別子、そして可能性としてクリップを定義するソース素材内の時間的範囲を用いて、ディジタル化ビデオまたはオーディオのような、ソース素材のクリップを引用する。識別子は、等価データ・ファイルのリストと共に用いて、ソース素材のファイル名を特定するものであれば、いずれの形式でもよい。インデックスを用いて、ソース内の時間範囲を、対応するファイル内のバイト範囲に変換することも可能である。このバイト範囲を、ファイルのセグメント・テーブルと共に用い、必要とされるセグメントおよびデータを検索する記憶装置を特定することができる。

【００８３】

図１４は、マルチメディア組成の一部を表わすために用いることができる、リスト構造の一例を示す。図１４に示す例では、数個のクリップ２６０があり、その各々が、２６２で示す、ソース識別子への引用、および２６４で示すように、ソース内の範囲を含む。一般に、時間組成において、メディアの各トラック毎に、このようなリストがあるとよい。組成を表わすために使用可能な種々のデータ構造がある。リスト構造に加えて、１９９３年１０月２８日に公開されたＰＣＴ公開出願ＷＯ９３／２１６３６には、更に複雑な構造が示されている。他のマルチメディア組成の代表例には、Ａｖｉｄ　Ｔｅｃｈｎｏｌｏｇｙ，Ｉｎｃ．からのＯｐｅｎ　Ｍｅｄｉａ　Ｆｒａｍｅｗｏｒｋ　Ｉｎｔｅｒｃｈａｎｇｅ　Ｓｐｅｃｉｆｉｃａｔｉｏｎ、Ｍｕｌｔｉｍｅｄｉａ　Ｔａｓｋ　Ｆｏｒｃｅからの　Ａｄｖａｎｃｅｄ　Ａｕｔｈｏｒｉｎｇ　Ｆｏｒｍａｔ（ＡＡＦ）、Ｍｉｃｒｏｓｏｆｔからの　ＤｉｒｅｃｔＳｈｏｗ、およびＰＣＴ公開ｗｏ９６／２６６００に示されているような、Ａｐｐｌｅ　ＣｏｍｐｕｔｅｒからのＢｅｎｔｏが含まれる。

【００８４】

先に述べ、マルチメディア・プログラムを表わすために用いられるデータ構造は、多数の形式のデータを用い、これらを同期させて表示することができる。最も一般的な例は、モーション・ビデオ（多くの場合、２つ以上のストリーム即ちトラック）および付随するオーディオ（多くの場合、４つ以上のストリーム即ちトラック）を含む、テレビジョン番組または映画の製作である。図１５に示すように、クライアント・コンピュータは、主メモリ内に割り当てたメモリ・バッファ２９４の対応するセット２９０を有することができる。各バッファは、「直列化」バッファとして実現することができる。言い換えると、クライアントは、記憶装置から受信したデータを、これら独立してアクセス可能な部分に挿入し、バッファ集合から順次読み出す。要求を数個の記憶装置に送ることができ、データは同じストリームに対して異なる時点で受信することができるので、バッファは、書き込み時には、連続的に満たされないこともあるが、読み出しは順次行われ表示される。図１５では、バッファ内の充填は、バッファ内におけるデータの存在によって示す。２９３および２９５で示すように、空のバッファはいずれも、いずれの時点でも充填することができる。しかしながら、各バッファ集合は、現リード・ロケーション２９１を有し、ここからデータを読み出し、かつ２９７に示すように、時間の進展に連れて前進する。これらのバッファの部分集合２９２，２９６を各データ・ストリームに割り当てることも可能である。

【００８５】

バッファ集合内の各バッファは、固定数のデータ・セグメントに対応するサイズを有し、セグメント・サイズとは、記憶装置上に格納したセグメントのサイズである。オーディオ・データ・ストリーム２９２当たり数個、例えば、４つのオーディオ・バッファがあるとした場合、各バッファは、数個、例えば、４つのセグメントを収容することができる。同様に、各ビデオ・ストリーム２９６は、数個、例えば、４つのバッファを有することができ、その各々は、数個、例えば、４つのセグメントを収容する。バッファの各々は、独立してアクセス可能な部分２９８に分割することができる。部分２９８のサイズは、ネットワークを通じた転送を予定しているデータ・パケットのサイズに対応する。

【００８６】

ビデオおよびオーディオ・データは異なるデータ・ファイルに格納することができ、しかも任意に組み合わせることができるので、クライアント側でこれら異なるストリームのデータ要求を効率的に管理すれば、一層の性能向上が得られる。例えば、クライアント・アプリケーションは、データを読み取ることができるストリームを識別し、次いで、データがあれば読み取るべきデータ量を決定することができる。この種のリード動作管理を行なうプロセスが、米国特許第５，０４５，９４０号に示されている。一般に、クライアントは、表示に利用可能なデータ量が最少のストリームはどれであるかについて判定を行なう。ある量のデータを効率的に読み出すために当該ストリームのために十分な量のバッファ空間がバッファ集合にある場合、そのデータを要求する。一般に、選択したストリームに対してメモリ内で利用可能な空間が十分に広く１ネットワーク伝送データ単位を保持できる場合、データは効率的に読み取られる。あるストリームのデータを要求すると判定した場合、各セグメントが格納されている記憶装置から選択した記憶装置から、データの各セグメントを要求する。

【００８７】

ある組成をデータ要求に変換し、当該データを表示するプロセスについて、図１６に関連付けてこれより説明する。記憶装置からどのファイルを要求するか知るために、クライアント・システム上で実行するアプリケーション・プログラムは、図１４に示すような組成を表わすデータ構造を、図１６のステップにおいてこれらのファイル内におけるファイル名および範囲に変換することができる。例えば、各ソース識別子およびそのソース内の範囲毎に、要求をアセット・マネージャに送ることができる。これに応答して、アセット・マネージャは、受信したソース識別子および範囲に対応する等価なメディアを収容するファイルのファイル名を戻すことができる。ファイルのセグメント・テーブルおよび利用可能な記憶装置のリストも、カタログ・マネージャとなることができる。

【００８８】

クライアントが特定のデータ・ストリームに対してデータ・セグメントを要求する場合、クライアントはステップ２７２において、要求したセグメントのため

に記憶装置を選択する。各セグメントをコピーすることによって冗長性を備えた一実施形態におけるこの選択について、図１７および図１８に関連付けて以下で更に詳しく説明する。一般に、キュー４８（図１）が最も短い記憶装置を選択すればよい。次にクライアントは、ステップ２７４ないし２７８において、選択した記憶装置から、そのセグメントに対するデータを読み出す。ステップ２７４は、プレリード・ステップとして理解するとよく、ここで、クライアントは記憶装置に要求を送り、所望のデータを、不揮発性ストレージからより高速な典型的に揮発性のストレージに読み出す。記憶装置に対する要求は、要求を行なってからクライアントにおいて要求データを受信しなければならないまで、即ち、期限までに要する時間がどれ位かかるかの指示を含むことができる。プレリード要求が受け入れられると、クライアントはステップ２７６において待機する。要求は記憶装置のキュー４８に置かれ、次いで期限を用いて以下に説明するように要求の優先順位を決める。ステップ２７８において、記憶装置のバッファにおいてデータが利用可能となった後、記憶装置からデータを転送する。このステップは、ネットワークの使用をスケジューリングしネットワーク利用効率を最大に高めることを含むとよい。受信したデータは、クライアントにおける適切なバッファに格納し、最終的にステップ２８０において処理し表示する。記憶装置においてセグメントが失われていた場合、冗長情報を用いてセグメントを再現する。

【００８９】

プレリード要求を開始する方法はいくつかあり、ステップ２７４における記憶装置の選択、およびステップ２７８におけるデータ転送を含む。例えば、マサチューセッツ州、ＴｅｗｋｓｂｕｒｙのＡｖｉｄ　Ｔｅｃｈｎｏｌｏｇｙ，Ｉｎｃ．からのＭｅｄｉａＣｏｍｐｏｓｅｒは、ユーザがクリップ数または時間量のいずれかをルック・アヘッド値（ｌｏｏｋ－ａｈｅａｄ　ｖａｌｕｅ）として設定させ、組成においてどの位先でアプリケーションがデータ・リード要求を開始すべきかを示す。テレビジョン放送説明のためのプログラム・スケジュールもこの目的のために使用可能である。このような情報は、記憶装置の選択およびプレリード要求を開始するために用いることができる。このようなプレリードは、１９９５年９月９日公開のヨーロッパ特許出願第０６７４４１４Ａ２号に示されるよ

うに、バッファ２９０（図１５）内にバッファ空間が得られない場合でも実行可能である。バッファ２９０（図１５）における利用可能な空間量を用いて、ステップ２７８（図１６）においてデータ転送を開始するか、あるいはプレリード（ステップ２７４）およびデータ転送（ステップ２７８）を開始することができる。

【００９０】

利用可能な記憶装置全てを洗いざらい探索する必要なく、クライアントが、最も短い要求キューを有する記憶装置の適切な推定を行なうことを可能にする１つのプロセスについて、図１７および図１８に関連付けてこれより説明する。最初に、クライアントは、ステップ３３０において、スレシホルドＥ１を有する要求を第１記憶装置に送る。スレシホルドＥ１は、要求に答えなければならない時間の推定値を示す。この推定値は、時間値、４というような記憶装置のディスク・キュー内の要求数、またはその他の尺度として表わすことができる。このスレシホルドの意味は、例えば、記憶装置が特定の時間制限の内に要求に答えることができる場合に、記憶装置によって要求が受け入れられなければならないということである。クライアントは、ステップ３３２において、記憶装置から回答を受信する。回答は、要求が受け入れられ記憶装置のディスク・キューに入力されたか、または要求が拒絶されたかを示す。これは、ステップ３３４において判定される。要求が受け入れられた場合、ステップ３３６において、記憶装置のバッファにおいてデータが利用可能となる時点の推定値がクライアントには与えられる。例えば、要求したセグメントのデータが既にバッファ内にある場合、記憶装置はデータが直ちに利用可能であることを示す。次いで、クライアントは、推定時間が経過して少し後に、データの転送を要求する時刻になるまで待てばよい。要求が拒絶された場合、ディスク・キューのエントリ数の実際のサイズのような、記憶装置が実際に要する可能性がある時間量の推定値が、記憶装置から戻される。ステップ３４０において、この実際の推定値を値Ｋに加算し、スレシホルドＥ２を得る。値Ｋは、ディスク・キューのエントリ数を表わす場合は、２となる。スレシホルドＥ１および値Ｋはユーザが定義可能である。ステップ３４２において、スレシホルドＥ２を示す要求を第２記憶装置に送る。次に、ステップ３４４に

おいて、クライアントは、ステップ３３２において受信した回答と同様に、回答を受信する。要求が受け入れらたことを回答が示すと３４６において判定した場合、クライアントは、ステップ３３６に示すように、第２記憶装置においてデータが利用可能となる時点の推定値を有し、その後クライアントは、データ転送をスケジュールするために待つことができる。それ以外の場合、ステップ３４８において、無条件要求、即ち、大きなスレシホルドを有する要求を第１記憶装置に送る。次に、ステップ３５０において承認を受信し、ステップ３３６において示したように、記憶装置のバッファにおいてデータが利用可能となる時点の推定値を示す。

【００９１】

一方、記憶装置は、要求を受信したときには、それがクライアントによって選択された第１記憶装置なのかまたは第２記憶装置なのかはわからない。逆に、記憶装置は、ステップ３６０に示すように、単に要求を受信するだけである。ステップ３６２において、要求内で示されるスレシホルドは、記憶装置自体の、クライアントが待たなければならない時間の推定値と比較される。例えば、記憶装置のディスク・キューのサイズを指定されたスレシホルドと比較する。供給内のスレシホルドが、記憶装置の推定値よりも大きい場合、ディスク・キューに要求を入力し、ステップ３６４において、記憶装置のバッファにおいてデータが利用可能となる時点の推定値を決定する。この推定値は、例えば、ディスク・アクセス速度、ディスク・キューの長さ、そして可能性として最新のパフォーマンスの移動平均に基づいて決定することができる。ステップ３６６において、承認をクライアントに送る。これは、記憶装置のバッファにおけるデータ可能性の推定時点を含む。それ以外の場合、ステップ３６８において拒絶を送り、ディスク・キューの実際の長さのような、この推定値を示す。

【００９２】

記憶装置は、記憶装置上のどのバッファに、どのセグメントがあるのか追跡することも可能である。セグメント・データは、記憶媒体からいずれかの空きバッファまたは最近最も使用されていないセグメントが占有するバッファに読み出すことができる。このように、あるセグメントに対するデータは、当該セグメント

を２回要求すれば、あるバッファにおいて直ちに利用可能とすることができる。

【００９３】

代替案として、クライアントは別の方法を用いて、データを検索する記憶装置を選択することも可能である。これについては、以下で論ずる。要求を送った後、クライアントは、記憶装置から、その要求が記憶装置のディスク・キュー内に入ったことを示す承認を受信する。記憶装置のバッファにおいてデータが利用可能となる時点の推定値を受信する代わりに、クライアントは、記憶装置が要求データを記憶装置の指定バッファ・メモリに読み込んだことを示すレディ信号を受信するまで待つこともできる。この待ち時間中、クライアントは、別のデータ・セグメントの要求を発行したり、データを表示したり、データを処理するというような、別のタスクを実行することができる。この代替案に伴う１つの問題は、クライアントが未要請メッセージ、即ち、記憶装置からのレディ信号を受け入れ、それに応答して、クライアントがコンテクストを変更し、メッセージを処理することである。クライアントは、他の動作を実行していて使用中の可能性もある。このプロセスは記憶装置のバッファにおいてデータが利用可能となる時点の更に精度が高い推定値を与えるが、コンテクストを変更し着信メッセージを処理する可能性のために、直ちにクライアントにおける複雑化を招くことになる。

【００９４】

セグメント・テーブルが各セグメントのコピーを追跡する場合に、あるファイルのためにセグメント・テーブルから記憶装置を選択する方法は、他にもいくつかある。例えば、クライアントがファイル・リード要求を行なっている場合、クライアントは、問題のファイルのために、「Ａ」リストまたは「Ｂ」リストのいずれかからランダムに抽出することができる。あるいは、クライアントは、その現在未処理の要求全て、即ち、送られたが未だ実行していない要求を見直し、ＡおよびＢリスト上の記憶装置の内、最も少ない未処理の要求を有する記憶装置を、ファイルに抽出することができる。この選択方法は、クライアントがそれ自体の未処理要求と競合する可能性を低下させることができ、要求を全ての記憶装置全体に一層等しく拡散するようになる。あるいは、未処理の要求を調べる代わりに、クライアントは、その最新要求の履歴、例えば最後の「ｎ」の要求を調べ、

次の要求のために、これまで最も使用されていないセグメントに対するＡリストおよびＢリストからいずれかの記憶装置を抽出することができる。この選択方法は、要求を全ての記憶装置全体に一層等しく拡散するようになり、更に特定の記憶装置に要求が集中するのを回避することができる。また、クライアントは、各記憶装置から、そのディスク・キューの長さの尺度を要求することもできる。クライアントは、最もディスク・キューが短い記憶装置に要求を発行することができる。別の可能性として、クライアントは、２つの記憶装置に要求を送り、最終的に一方のみからデータを受信することもできる。この方法をローカル・エリア・ネットワーク上で用いると、クライアントは未使用の要求を取り消すことができる。ワイド・エリア・ネットワーク上では、最終的に選択された記憶装置は、他の記憶装置において未使用の要求を取り消すことができる。

【００９５】

記憶装置は、多数のアプリケーションから多数の要求を受信する可能性もある。多数のアプリケーションからの要求を管理し、最も重要な要求を最初に処理することを保証するために、各記憶装置毎にキュー４８（図１）を維持する。キューは、システムの複雑度に応じて、いくつかの部分に維持することも可能である。即ち、記憶装置は、ディスク・アクセスのためおよびネットワーク転送のために異なるキューを維持することができる。キューは、例えば、再生から放送のために特定の時間期限があるデータを用いる時間厳守アプリケーションからの要求を、キャプチャ・システム、オーサリング・ツールまたはサービスおよび保守アプリケーションのようなその他のアプリケーションからの要求から、分離することも可能である。更に、格納要求は、オーサリング・ツールからの要求、ならびにサービスおよび保守プログラムからの要求から分けることができる。オーサリング・ツールからの要求は、更にサービスおよび保守要求から分けることができる。

【００９６】

図１９は、ディスク・キュー３００およびネットワーク・キュー３２０を利用するキュー４８の一実施形態を示す。ディスク・キューは４つのサブキュー３０２，３０４，３０６および３０８を有し、それぞれ、再生、キャプチャ、オーＡ

リング、ならびにサービスおよび保守クライアント・プログラムに１つずつである。同様に、ネットワーク・キュー３２０も４つのサブキュー３２２，３２４，３２６および３２８を有する。各キューは、１つ以上のエントリ３１０を含み、その各々は、要求を行なったクライアントおよび要求された動作を示す要求フィールド３１２、要求の優先度を示す優先度フィールド３１４、ならびに要求に関連するバッファを示すバッファ・フィールド３１６を含む。要求の優先度の指示は、デッドライン、タイム・スタンプ、クライアントにおいて利用可能なメモリ量の指示、またはクライアントにおいて現在利用可能なデータ量の指示とすることができる。記憶装置に優先度スケジューリング機構があれば、使用する優先度スタンプの種類を指示することができよう。

【００９７】

優先度値は多くの方法で生成することができる。オーサリングまたは再生システムに対する優先度値は、一般に、アプリケーションが要求データを受信しなければならない時点の尺度である。例えば、リード動作では、アプリケーションは、データがなくなる目に、再生に利用できるデータがどの位あるかについて（ミリ秒単位またはフレーム単位またはバイト単位で）報告することができる。キャプチャ・システムに対する優先度指示は、一般に、クライアントがそのバッファから記憶装置にデータを転送しなければならない時点の尺度である。例えば、ライト動作では、アプリケーションは、バッファが溢れる前に満たすことができる空のバッファ空間がどれ位あるかについて（ミリ秒単位、フレーム単位またはバイト単位で）報告することができる。ミリ秒を尺度単位として用いた場合、システムは、絶対的な時間クロックを有し、これをキュー４９における要求の順番を決定するための基礎として用いることができ、全てのアプリケーションおよび記憶装置をこの絶対時間クロックに同期させることができる。この同期が実用的でない場合、アプリケーションは、当該アプリケーションに関係し、要求を行って渡されてから、要求データがクライアントによって受信されるまでどれ位の時間を要するかを示す時間を用いることも可能である。通信レイテンシが低いと仮定すると、記憶装置は、この相対時間を、記憶装置と一貫性のある絶対時間に変換することも可能である。

【００９８】

記憶装置は、そのサブキュー３０１ないし３０８において、要求をその優先度の順に処理する。即ち、最初に最も優先度が高いキューにある要求をそれらの優先度値の順に処理し、次いで、次第に優先度が低くなるキューにおける要求を処理する。各要求毎に、記憶装置は、ディスクと要求によって示されるバッファとの間でデータを転送する。リード要求では、要求を処理した後、ディスク・キューからネットワーク・キューに要求を移管する。ライト要求では、ライト動作の完了に成功した後、ディスク・キューから要求を除去する。

【００９９】

以下で更に詳しく説明する一実施形態では、記憶装置はネットワーク・キューを用いて、これらの転送をスケジュールするプロセスにおいて、ネットワーク転送の優先順位を決定する。この実施形態では、クライアントは、ネットワークを通じてデータの転送を要求する。記憶装置がこのような要求を２つほぼ同時に受信した場合、記憶装置は、そのネットワーク・キューにおいて優先度が高い方の要求を処理する。リード要求では、要求を処理した後、ネットワーク・キューから要求を除去する。ライト要求では、転送完了が成功した後、要求をネットワーク・キューからディスク・キューに移管する。優先度は、空きバッファの可用性に依存する。ネットワーク・キュー内にある要求を処理する時間が過ぎた場合、要求を欠落させ、クライアントはもはや動作していないか、またはネットワーク転送を時間内に要求しなかったことを示す。

【０１００】

記憶装置とクライアントとのコンピュータ・ネットワークを通じたデータ転送をスケジュールし、効率を向上させることも可能である。即ち、データ転送のスケジューリングにより、コンピュータ・ネットワークの帯域幅利用度を高める。このようなネットワーク使用のスケジューリングは、特にクライアントとスイッチとの間のリンクの帯域幅が、記憶装置とスイッチとの間のリンクの帯域幅と同程度の大きさである場合に実行するとよい。即ち、記憶装置がデータを送り、クライアントがそれらそれぞれのネットワーク接続のリンク速度でデータを受信する場合、データはネットワーク・スイッチにおいて蓄積したり、他の大幅な遅れ

が生ずる可能性は低くなる。

【０１０１】

このようなネットワークの利用度を高めるために、いずれの所与の時点においても、各クライアントに１つの記憶装置のみからデータを受信させ、各記憶装置に１つのライアンとのみにデータを送らせる機構を備えるとよい。例えば、各クライアントがトークンを１つのみ有するようにするとよい。クライアントは、このトークンを１つの記憶装置のみに送り、選択したセグメントに対するデータの転送を要求する。トークンは、クライアントがデータを受信しなければならないデッドライン、即ち、優先度の尺度および指定セグメントを示すことができる。各記憶装置は一度に、それがトークンを受信した１つのクライアントだけにデータを送る。記憶装置は、一度に１つのトークンのみを受信する。データを転送した後、記憶装置はトークンも返す。

【０１０２】

図２０および図２１に関連付けて他のネットワーク・スケジューリング・プロセスについてこれより説明する。このプロセスは、同様の結果を与えるが、トークンを使用しない。むしろ、クライアントが記憶装置との通信チャネルを要求し、セグメント、およびクライアントが転送が行われるのを待つ時間量Ｅ３を指定する。また、クライアントは、クライアントがデータを受信しなければならない新たな時間期限をセグメントに対して指定することもできる。

【０１０３】

これより図２０を参照して、ネットワークを通じてデータを転送するクライアント・プロセスについて説明する。組成の再生中のいずれかの時点において、各バッファは、それに関連するデータ・セグメント、および連続再生のためにバッファ内においてデータが利用可能でなければならない時点を有する。当技術分野では公知であるが、アプリケーションは、再生プロセスの間、バッファの各々をセグメントに関連付ける。図１７および図１８に関連付けて先に示したように、クライアントが準備した各セグメントには、記憶装置においてデータが利用可能となる推定時間が関連付けられている。したがって、クライアントは、バッファをそれらの時間期限によって、および要求データが記憶装置のバッファ内におい

て利用可能であることが求められているか否かによって順序付けることができる。この順序付けは、クライアントが、ステップ５００においてデータを転送するために次のバッファを選択するために用いることができる。クライアントは、ステップ５０２において、記憶装置との通信チャネルを要求し、待ち時間Ｅ３を指定する。クライアントがデータを早急に必要としない場合、またはクライアントが他の動作をより効率的に実行できる場合、この値Ｅ３は、例えば、１００ミリ秒と短くてよい。この値Ｅ３は、クライアントが早急にデータを必要とする場合には、例えば、そのバッファの１つでデータが無くならないように長くすることも可能である。ステップ５０４において、クライアントは記憶装置から回答を受信する。記憶装置が要求を拒絶したとステップ５０６において判定した場合、ステップ５０８において改訂推定時間をメッセージと共に受信する。ステップ５１０において、この改訂推定時間を用いて、バッファを選択するバッファ・リストを更新することができる。処理はステップ５００に戻り、別のバッファを選択する。セグメントが、以前に選択したセグメントと同じ記憶装置上にあるバッファは、恐らく選択すべきでない。記憶装置が別の方法で要求を受け入れた場合、最終的にステップ５１８でデータを受信する。

【０１０４】

記憶装置の観点からの処理について、図２１に関連付けてこれより説明する。ステップ５２０において、記憶装置は、待ち時間Ｅ３を示す要求をクライアントから受信する。この記憶装置のバッファにではデータが未だ利用可能でないとステップ５２２において判定した場合、ステップ５２４において記憶装置は要求を拒絶し、改訂推定時間を計算し、これをクライアントに送る。逆に、データが利用可能であり、記憶装置のネットワーク接続が使用中でないとステップ５２６において判定した場合、ステップ５２８においてクライアントは「アクティブ・クライアント」となり、記憶装置によって通信チャネルが付与され、データの転送が可能となる。記憶装置のネットワーク接続が他のクライアントにデータを転送するために稼働している場合、記憶装置は「待機クライアント」からの要求を維持し、「アクティブ・クライアント」に対するデータ転送が完了した後に、「待機クライアント」にデータを転送する。現クライアントが「待機クライアント」

か否かについて判定を行なうために、ステップ５３０において、記憶装置は、ネットワーク・キューにおいてデッドラインが速い要求の数に、各要求毎のネットワーク伝送時間を乗算した値に基づいて、転送を行なうことができるまでの時間を推定する。計算した利用可能推定時間が待ち時間Ｅ３よりも大きいとステップ５３２で判定した場合、クライアントはそんなに長く待ちたくないことを示し、ステップ５２４において要求を拒絶する。また、この要求の指定優先度が、いずれの現待機クライアントの優先度よりも低いとステップ５３４において判定した場合、ステップ５２４において要求を拒絶する。それ以外の場合、ステップ５３６において、いずれの現待機クライアントからの要求も拒絶し、この新たなクライアントを現待機クライアントとして指定する。アクティブ・クライアントへの転送が完了すると、待機クライアントがアクティブ・クライアントとなり、データを転送する。

【０１０５】

クライアントから記憶装置にデータを転送するためには、同様のプロセスを用いて、ネットワーク転送をスケジュールし、記憶装置内のバッファから不揮発性ストレージにデータを転送することができる。クライアントの観点から、このプロセスについて図２２に関連付けてこれより説明する。このプロセスは、図３におけるステップ１２４および１２６を実行するために用いることができる。

【０１０６】

クライアントがデータをそのバッファ集合内の任意の点に入力することができるリード・プロセスとは異なり、記憶装置に転送するデータは、典型的に、キャプチャ・システムが使用するバッファ集合からのリード・ポインタから来る。

【０１０７】

キャプチャ・システムは、典型的に、１つ以上のビデオ情報ストリーム、および１つ以上のオーディオ情報ストリームを生成する。したがって、キャプチャ・システムは、ストリーム内の空きバッファ空間量に応じて、データ・ストリームの１つを選択し、捕獲したデータを受信することができる。この選択したストリームの現リード・ポインタにおけるバッファを、ステップ６００において選択する。次に、ステップ６０２においてライト要求を記憶装置に送る。要求は、セグ

メントの識別子、時間期限またはその他の優先度値、およびクライアントが待とうとする時間量を示すスレシホルドＥ４を含む。時間期限は、記憶装置がネットワーク転送要求の優先度を決定するために用いる。スレシホルドＥ４は、先に論じたスレシホルドＥ３と同様、クライアントが用い、それ自体の動作を効率的にスケジュールすることを可能にする。クライアントは、要求を記憶装置に送った後、最終的にステップ６０４において回答を受信する。ライト要求が拒絶されたことを要求が示すとステップ６０６において判断した場合、回答は、ステップ６０７においてデータを受信するために記憶装置が利用可能となるまでの推定時間を含む。この推定時間は、クライアントが他の動作をスケジュールするために用いることができる。記憶装置がデータを書き込む要求を受け入れた場合、ステップ６０８において、クライアントは、データのセグメントの一部を記憶装置に送る。ステップ６１０において、ライト要求が成功したか否かを示す回答をステップ６１０で受信することができる。これは、ステップ６１２において分析する。ステップ６１４において、不良の場合には、復元プロセスを伴う場合もある。それ以外の場合、ステップ６１６に示すように、プロセスは完了する。

【０１０８】

記憶装置の観点から、記憶装置はステップ６２０においてクライアントからのライト要求を受信する。要求は、時間期限またはその他の優先度スタンプを含み、これを用いて要求をネットワーク・キューに入力する。次に、ステップ６２２において、記憶装置はデータを受信するためにバッファが利用可能か否かについて判定を行なう。利用可能なバッファがないというありそうもない場合には、ステップ６２４において要求が拒絶される場合もある。それ以外の場合に、ステップ６２６において要求をネットワーク・キューに入力し、データ、その優先度スタンプ、および転送に関するその他の情報を受信するためにバッファを割り当てたことを示す。次に、記憶装置は、ステップ６２８において、ネットワーク接続が使用中か否かについて判定を行なう。ネットワーク接続が使用中でない場合、記憶装置はステップ６３０において要求を受け入れ、その旨のメッセージをクライアントに送る。すると、クライアントはデータを転送し、これがステップ６３２において記憶装置によって受信され、指定のバッファに入力される。指定のバ

ッファが現在満杯であるとステップ６３４において判定した場合、ステップ６３６において、適切な優先度スタンプと共にバッファをディスク・キューに入力する。記憶装置のディスク・キューの処理によって、最終的にデータはバッファから永続ストレージに転送される。それ以外の場合、ステップ６３８に示すように、クライアントがバッファを満たすのに十分なデータを送るまで、記憶装置は待機する。

【０１０９】

記憶装置のネットワーク接続が使用中であるとステップ６２８において判定した場合、記憶装置はステップ６４０において、記憶装置のネットワーク接続が利用可能となるまでの推定時間を計算する。この計算時間が、ステップ６４２に判定した指定の待ち時間Ｅ４よりも大きい場合、ステップ６４３において、記憶装置の利用可能時間の推定値によって、要求を拒絶する。記憶装置が、クライアントによって示された待ち時間Ｅ４以内にデータを転送することができる場合、ステップ６４４において、記憶装置は、要求の優先度を、現在待機中のあらゆるクライアントの要求の優先度と比較する。この要求の優先度が、現在待機中のクライアントの要求よりも低い場合、この要求を拒絶する。それ以外の場合、現在待機中のクライアントからの要求を拒絶し、この新たな要求を次の要求とし、ステップ６４６において処理する。

【０１１０】

冗長情報を２つ以上のセグメントから作成する場合に用いる追加実施形態について、これより図２４および図２５に関連付けて説明する。

【０１１１】

まず図２４を参照して、数個の記憶装置にわたってランダムに分散して冗長情報と共にデータ・セグメントを格納するプロセス例について詳細に説明する。このプロセスは、図３に関連付けて先に説明したプロセスと全体的に類似している。最初に、ステップ７００において、キャプチャ・システムはセグメント・テーブル９０Ｂ（図２Ｂ）を作成する。典型的に、各画像を、捕獲するデータ・ストリームにおけるオフセットにマップする画像インデックスも作成する。インデックス化した画像は、例えば、ビデオのフィールドまたはフレームに対応すること

ができる。インデックスは、オーディオのような他の種類のデータに対する、時間期間のような、別のサンプル境界を基準とすることができる。また、キャプチャ・システムは、前述のような、利用可能な記憶装置のリストも得る。また、キャプチャ・システムは、冗長セット・サイズの指示を、利用可能な記憶装置のリストに基づいて自動的に、またはユーザから受け取る。一般に、冗長セット・サイズは、利用可能な記憶装置の数未満でなければならず、非常に小さい部分集合であってもよい。また、カウンタを用いて、どのセグメントが所与の冗長セット内にあるか追跡する。このカウンタは、ステップ７００においてゼロにリセットする。排他的ＯＲメモリも用い、全て二進無アサート値、例えば、「０」にリセットする。

【０１１２】

次に、ステップ７２０で、データ・セグメントをキャプチャ・システムにおいて作成する。このセグメントに適切なサイズについては、図３の説明との関係で先に論じた。また、ステップ７２０ではカウンタの増分も行なう。

【０１１３】

ステップ７２２において、現セグメントは、排他的ＯＲメモリに既に格納したあらゆるセグメントの排他的ＯＲとしてローカルに格納する。ステップ７２４において、記憶装置をこのセグメントのために選択する。セグメントのための記憶装置の選択は、ランダムまたは疑似ランダムである。この選択は、以前のいずれの冗長セットに対して行なった選択にも独立とすることができる。しかしながら、選択は、冗長セット内の各セグメントが異なる記憶装置上に格納されていることを保証しなければならない。図３の説明との関係で先に論じたように、各ファイルは、利用可能な記憶装置の部分集合のみを使用することができる。

【０１１４】

セグメントに記憶装置を選択した後、ステップ７２６においてこのセグメントを記憶装置に送り格納する。次に、ステップ７２８において、キャプチャ・システムは、記憶装置がセグメントの格納完了を承認するまで待つ。データを捕獲しながらリアル・タイムで格納しなければならない場合、ステップ７２６におけるデータ転送は、先に論じたようにリード動作と同様、２段階で行なえばよい。デ

ータを記憶装置に格納するのに成功した後、ステップ７３０においてキャプチャ・システムはセグメント・テーブル９０Ｂを更新する。

【０１１５】

カウンタが現在冗長セットのサイズに等しいとステップ７３２において判定した場合、ローカル排他的ＯＲメモリの内容は冗長情報である。この場合、この冗長情報を記憶装置上に格納する。即ち、ステップ７３４においてカウンタをリセットする。ステップ７３６において、冗長情報のために記憶装置を選択する。ステップ７３８において、冗長情報を選択した記憶装置に送る。次いで、ステップ７４０において、キャプチャ・システムは格納成功の承認を待つ。ステップ７４２において、セグメント・テーブルを更新することができる。

【０１１６】

捕獲が完了したとステップ７４４において判断した場合、プロセスは終了する。この時点で、ステップ７４５において、排他的ＯＲメモリに格納されているあらゆる冗長情報は、ステップ７３４ないし７４２と同様の手順を用いて、記憶装置に格納しなければならない。次に、ステップ７４６において、更新したセグメント・テーブルをカタログ・マネージャに送る。ステップ７３２においてカウンタが冗長セットのサイズに等しくない場合、また捕獲が完了していないとステップ７４４において判断した場合、プロセスは継続し、ステップ７２０においてデータの次のセグメントを作成し、カウンタを増分する。

【０１１７】

図５に関連付けて先に論じたように、冗長情報は、記憶装置の１つが故障した場合に、データを復元することを可能にする。図２５は、冗長情報が、２つ以上のセグメントを収容する冗長セットに基づく場合に、このような故障回復を行なうプロセスを示す。図５に示すように、復元するファイルをステップ７５０において選択する。ステップ７５２において、そのファイルの失われたセグメントを全て特定する。次に、ステップ７５４において、失われたセグメントを収容する冗長セットを読み取る。このステップは、当該セットにおけるセグメントの排他的ＯＲによって作成したセットに対する冗長情報を読み出し、次いで冗長セットの残りのセグメントを読み出すことからなる。次に、ステップ７５６において残

りのセグメントおよび冗長情報の排他的ＯＲを計算し、失われたセグメントを再構成する。図５のステップ２０４と同様に、次に、ステップ７５８において、失われたセグメントを再構成したもの毎に記憶装置を選択する。選択した記憶装置に、失われたセグメントを再構成したものを格納する。ステップ７６０において、格納動作の完了成功時に、セグメント・テーブルを更新する。次に、ステップ７６２において、更新したセグメント・テーブルをカタログ・マネージャに送る。

【０１１８】

また、一種の冗長情報、例えば、セグメントのコピーを有するファイルを、他の種類の冗長情報、例えば、２つ以上のセグメントの排他的ＯＲに変換することも可能である。例えば、図６に示すプロセスを用いて、データの追加コピーを作成することができる。このプロセスを完了した後、他の形態の冗長データ（セグメントの排他的ＯＲの結果）を削除することができる。同様に、図２４に示すプロセスは、格納したデータと共に用いて、冗長情報の排他的ＯＲを作成することも可能である。このような情報を作成した後、図７に示したプロセスを用いてデータの余分なコピーを削除することができる。ファイルが冗長情報を有する形態は、ファイル毎に異なってもよく、更に、例えば、ファイルに関連する優先度に基づくことも可能であり、冗長情報の形態の指示をカタログ・マネージャに格納することも可能である。

【０１１９】

ネットワークを通じたデータ転送をスケジュールすることにより、そして記憶装置上の負荷を分散し、冗長情報と共にランダムに分散したデータ・セグメントに対するアクセスを選択することによって、このシステムは、多数のアプリケーションと多数の記憶装置との間で双方向に多数のデータ・ストリームを非常にスケーラブルにかつ信頼性高く、効率的に転送することが可能となる。これは、分散型マルチメディア製品にとっては特に有効である。

【０１２０】

このようなコンピュータ・ネットワークを用いて実現可能なアプリケーションの１つは、多数のストリームを他の外部ディジタル効果システムに送りかつ戻し

、ライブ製作において共通に用いる機能である。これらのシステムは複雑でコスト高となる場合がある。殆どのディスクを用いたノンリニア・ビデオ編集システムは、効果リターン・ストリームを同時に記録しながら多数の再生ストリームを維持することができない、ディスク・サブシステムおよびバス・アーキテクチャを有し、オンライン環境で用いる可能性を制限している。このシステムを用いれば、数種類のストリームを効果システムに送ることができ、効果システムは効果データ・ストリームを出力し、多数の記憶装置上に格納する。数種類のストリームは、多数のカメラ・ソース、またはデュアル・ディジタル・ビデオ効果のレイヤとすることができる。

【０１２１】

また多数の記憶装置を有し、データを１つのクライアントに供給し、他のいずれの記憶装置よりも高い帯域幅を有する高帯域幅データ・ストリームを必要とするクライアントを満足させることも可能である。例えば、２０個の記憶装置の各々がスイッチに対して１０ＭＢ／ｓリンクを有し、クライアントがスイッチに対して２００ＭＢ／ｓリンクを有する場合、クライアントは２０個の記憶装置から同時に２００ＭＢ／ｓを読み取ることができ、例えば、高品位テレビジョン（ＨＤＴＶ）のデータ・ストリーム転送が可能となる。

【０１２２】

これまでに概略的に説明した手順を用いると、記憶装置およびクライアントは、ローカル情報を用い、中央のコンフィギュレーション管理または制御なく動作する。記憶装置は、システムを停止する必要なく、動作の間にシステムに追加することができる。記憶装置は単に動作を開始し、クライアントにその可用性を知らせ、次いでアクセス要求に対して応答するプロセスを確立する。この拡張性は、システムの機能および信頼性を補完するものである。

【０１２３】

以上いくつかの実施形態について説明したが、これまでに述べたことは単なる例示であって限定ではなく、一例として提示したに過ぎないことは、当業者には明白であろう。多数の変更や別の実施形態は当業者の範囲内のことであり、添付した特許請求の範囲およびその均等物に該当するものと見なすこととする。

【図面の簡単な説明】

【図１】

図１Ａは、コンピュータ・システムの一例のブロック図である。図１Ｂは、図１Ａのシステムの別の実施形態のブロック図である。

【図２】

図２Ａは、図１Ａにおける記憶装置４２にデータ・セグメントをマッピングするデータ構造を示す。図２Ｂは、図１Ｂにおけるデータ記憶装置４２のセグメントをマッピングするデータ構造を示す。

【図３】

一実施形態において、どのようにしてデータを捕獲し数個の記憶装置間で分散するのかについて記載したフローチャートである。

【図４】

一実施形態において、どのようにして記憶装置はデータ格納要求を処理するのかについて記載したフローチャートである。

【図５】

記憶装置が利用できなくなった場合に、どのようにして故障回復を実行するかについて記載したフローチャートである。

【図６】

追加データ・コピーをどのようにして作成することができるかについて記載したフローチャートである。

【図７】

データ・コピーをどのようにして削除することができるかについて記載したフローチャートである。

【図８】

記憶装置をどのようにしてシステムから除去することができるかについて記載したフローチャートである。

【図９】

どのようにしてデータを保管またはバックアップとしてコピーすることができるかについて記載したフローチャートである。

【図１０】

カタログ・マネージャに記憶装置の可用性を通知するための、記憶装置上での処理の状態図である。

【図１１】

カタログ・マネージャが維持することができる記憶装置のリストを示す。

【図１２】

どのようにしてカタログ・マネージャは記憶装置を監視することができるかについて示す状態図である。

【図１３】

メディア・データ・ファイルの同等性を追跡するためのテーブルを示す。

【図１４】

数個のクリップから成るモーション・ビデオ・シーケンスを表わすリスト構造を示す。

【図１５】

２つのモーション・ビデオ・データ・ストリーム、および４つの関連するオーディオ・データ・ストリームの再生にクライアントにおいて対応するためのバッファ・メモリの構造を示す。

【図１６】

クライアントはいかにしてマルチメディア組成を処理し、選択した記憶装置からのデータに対する要求にするのかについて記載したフローチャートである。

【図１７】

一実施形態において、クライアントはいかにして記憶装置に一次ストレージからバッファにデータ転送を要求するかについて記載したフローチャートである。

【図１８】

記憶装置はいかにして図１７におけるクライアントからの要求に回答するのかについて記載したフローチャートである。

【図１９】

データに対するディスク・アクセス要求に優先順位を付けるためのディスク・キューの一例、およびネットワーク・データ転送要求に優先順位を付けるための

ネットワーク・キューを示す。

【図２０】

一実施形態において、クライアントがどのようにして記憶装置にネットワークを通じたデータ転送を要求するかについて記載したフローチャートである。

【図２１】

一実施形態において、記憶装置がどのようにして多数のクライアントからのデータ転送要求を処理するかについて記載するフローチャートである。

【図２２】

クライアントがデータを当該クライアントから記憶装置に転送する際に実行するネットワーク・スケジューリング・プロセスの一実施形態について記載するフロー・チャートである。

【図２３】

記憶装置がデータをクライアントから当該記憶装置に転送する際に実行するネットワーク・スケジューリング・プロセスの一実施形態について記載するフロー・チャートである。

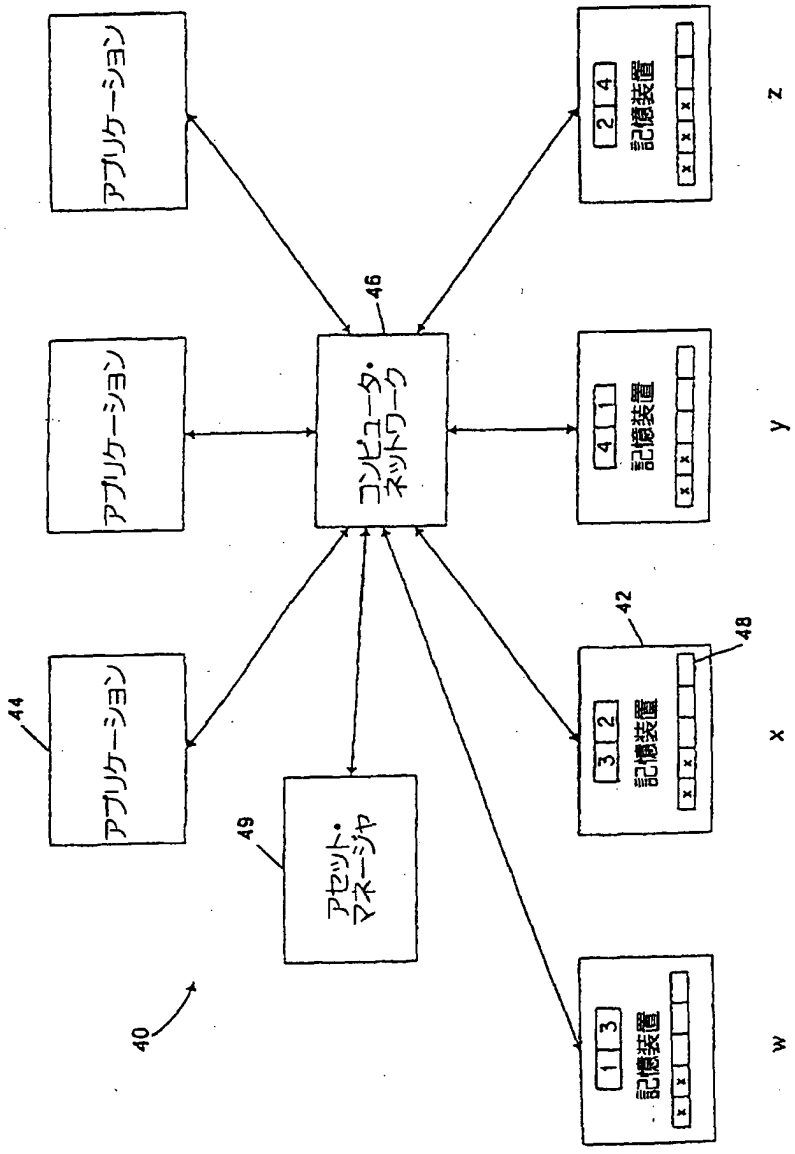
【図２４】

別の実施形態において、どのようにしてデータを捕獲し、数個の記憶装置間で分散することができるのかについて記載したフロー・チャートである。

【図２５】

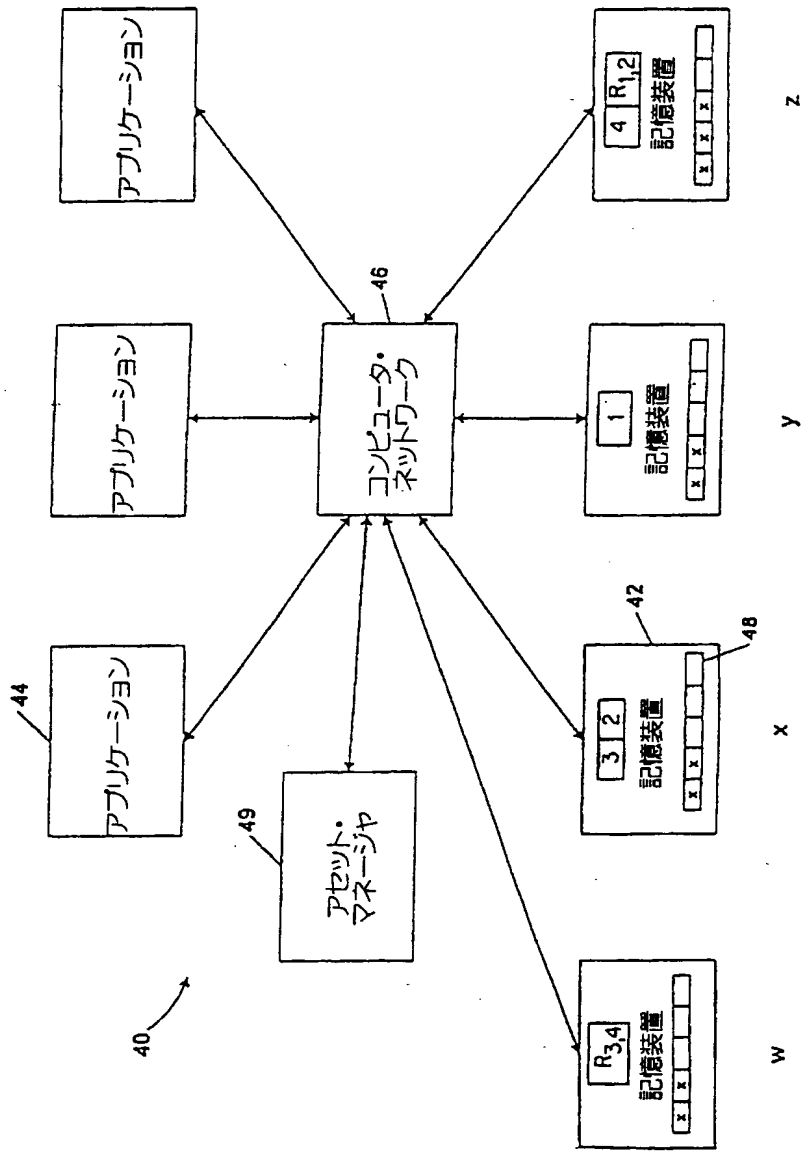
別の実施形態において、記憶装置が使用できなくなった場合に、どのようにして故障回復を実行することができるかについて記載したフロー・チャートである。

【図１Ａ】


アプリケーション
アプリケーション
アプリケーション
コンピュータ・ネットワーク
アセット・マネージャ
記憶装置
記憶装置
記憶装置
記憶装置
40
42
44
46
48
49
w
x
y
z

【FIG. 1A】

【図１Ｂ】


アプリケーション
アプリケーション
アプリケーション
コンピュータ・ネットワーク
46
44
49
アセット・マネージャ
40
42
48
4 R1,2 記憶装置
1 記憶装置
3 2 記憶装置
R3,4 記憶装置
z
y
x
w

【FIG. 1B】

【図２Ａ】

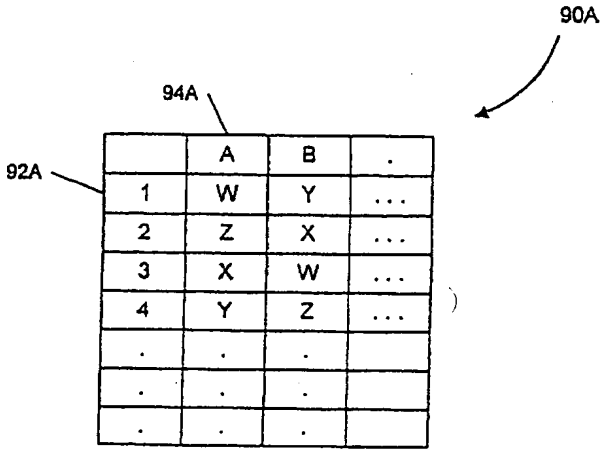


FIG. 2A

【図２Ｂ】

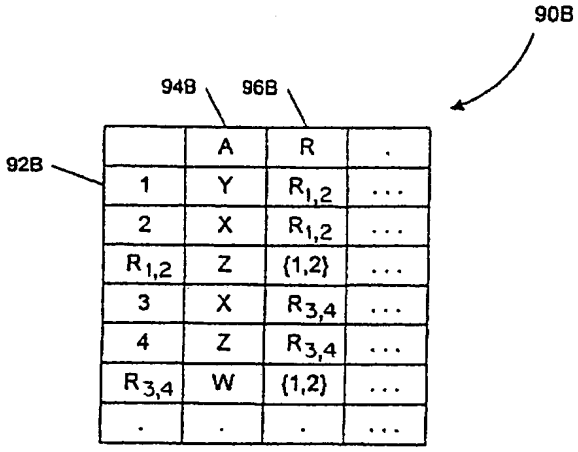
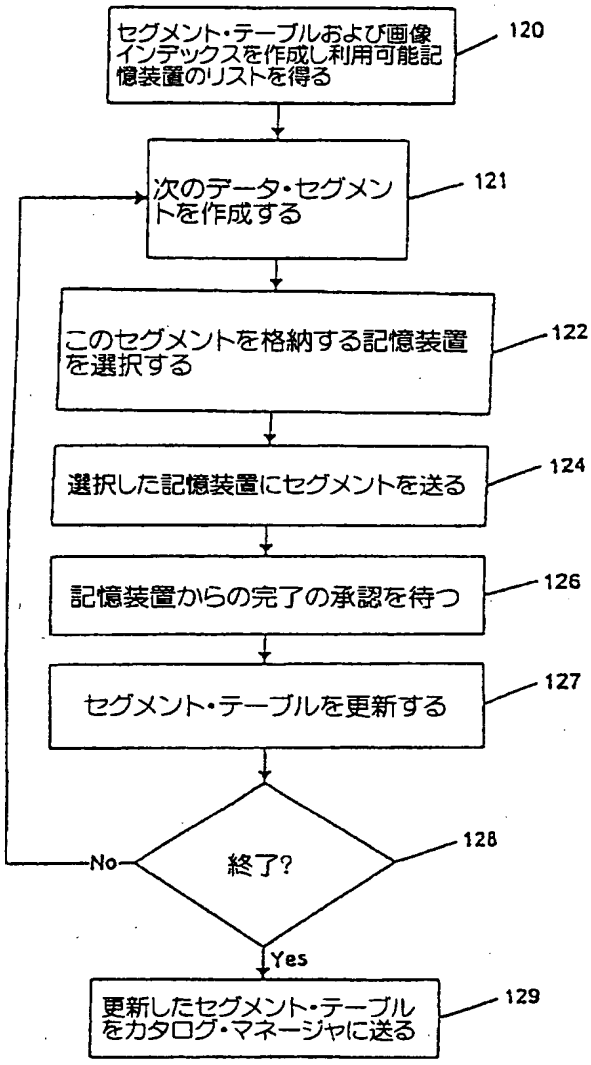


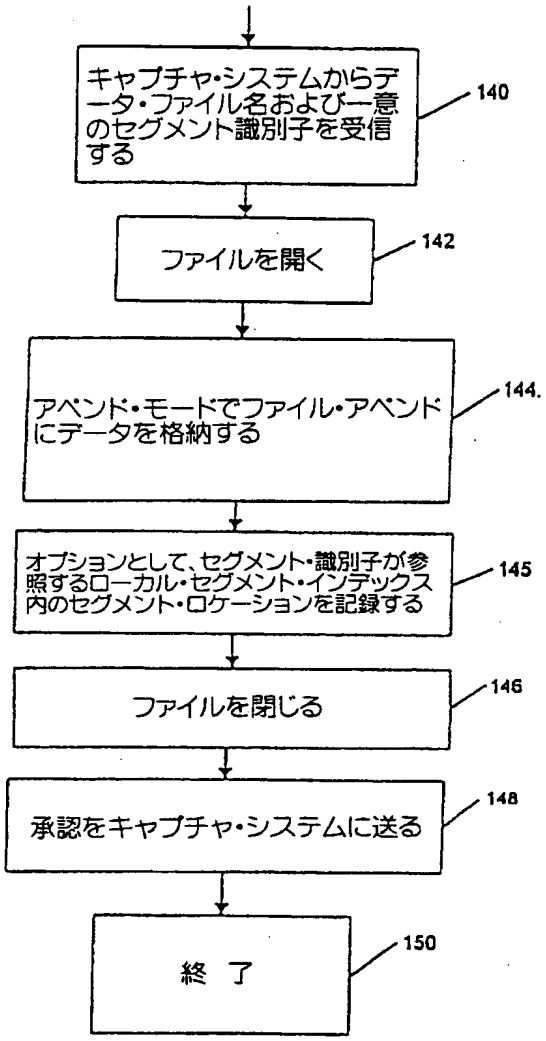
FIG. 2B

【図３】


セグメント・テーブルおよび画像インデックスを作成し利用可能記憶装置のリストを得る
120
次のデータ・セグメントを作成する
121
このセグメントを格納する記憶装置を選択する
122
選択した記憶装置にセグメントを送る
124
記憶装置からの完了の承認を待つ
126
セグメント・テーブルを更新する
127
終了?
128
No
Yes
更新したセグメント・テーブルをカタログ・マネージャに送る
129

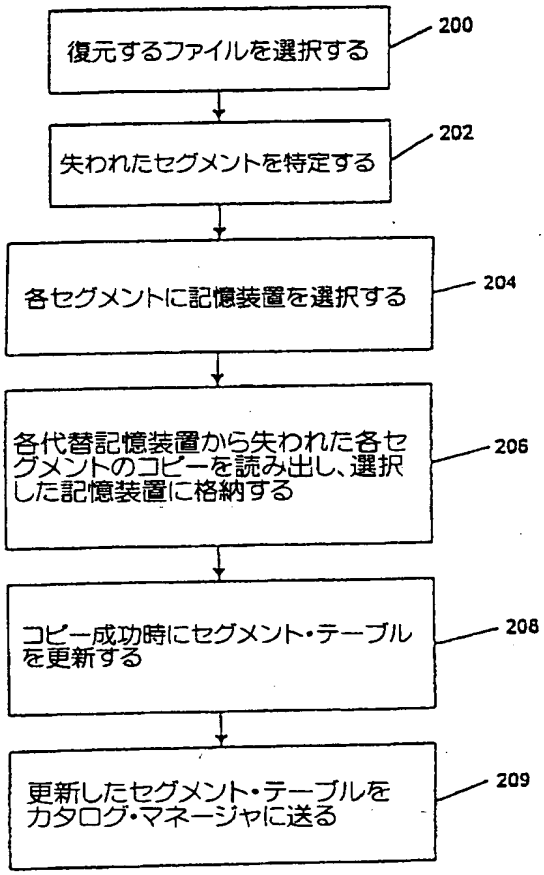
【FIG. 3】

【図４】


キャプチャ・システムからデータ・ファイル名および一意のセグメント識別子を受信する
140
ファイルを開く
142
アペンド・モードでファイル・アペンドにデータを格納する
144
オプションとして、セグメント・識別子が参照するローカル・セグメント・インデックス内のセグメント・ロケーションを記録する
145
ファイルを閉じる
146
承認をキャプチャ・システムに送る
148
終 了
150

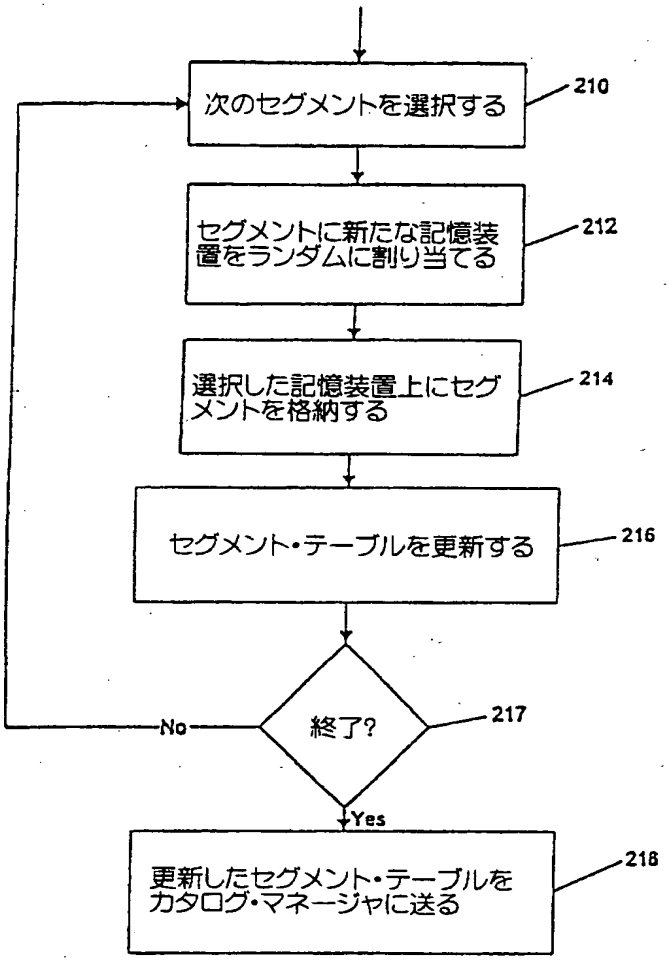
【FIG. 4】

【図５】


復元するファイルを選択する
200
失われたセグメントを特定する
202
各セグメントに記憶装置を選択する
204
各代替記憶装置から失われた各セグメントのコピーを読み出し、選択した記憶装置に格納する
206
コピー成功時にセグメント・テーブルを更新する
208
更新したセグメント・テーブルをカタログ・マネージャに送る
209

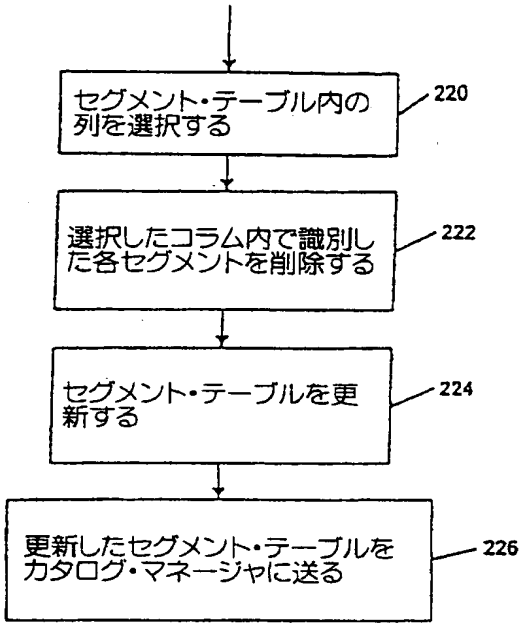
【FIG. 5】

【図６】


次のセグメントを選択する
210
セグメントに新たな記憶装置をランダムに割り当てる
212
選択した記憶装置上にセグメントを格納する
214
セグメント・テーブルを更新する
216
終了?
217
No
Yes
更新したセグメント・テーブルをカタログ・マネージャに送る
218

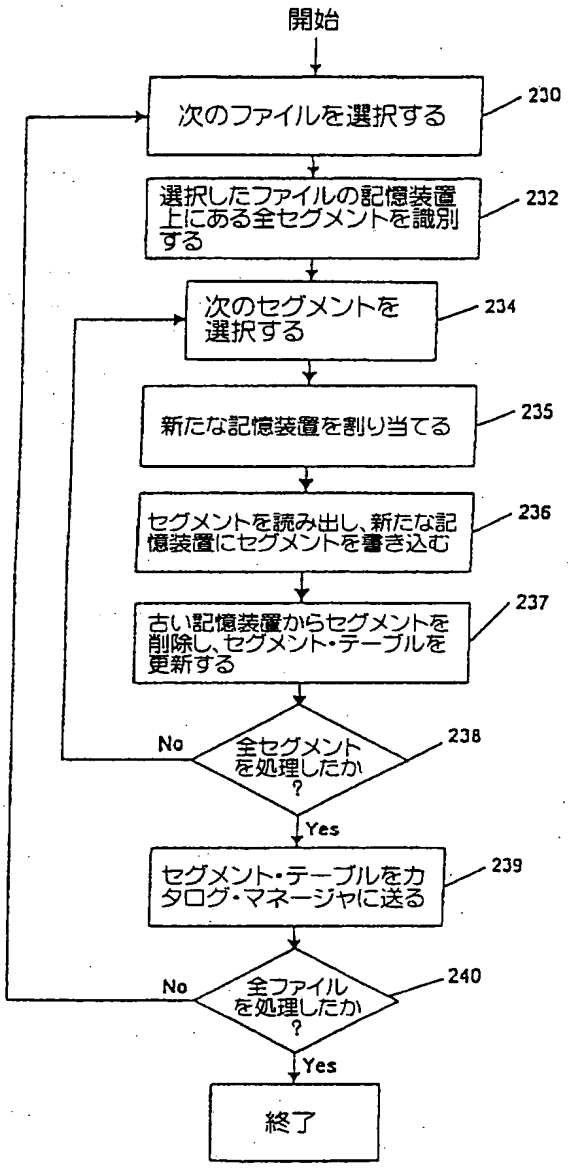
【FIG. 6】

【図７】


セグメント・テーブル内の列を選択する
220
選択したコラム内で識別した各セグメントを削除する
222
セグメント・テーブルを更新する
224
更新したセグメント・テーブルをカタログ・マネージャに送る
226

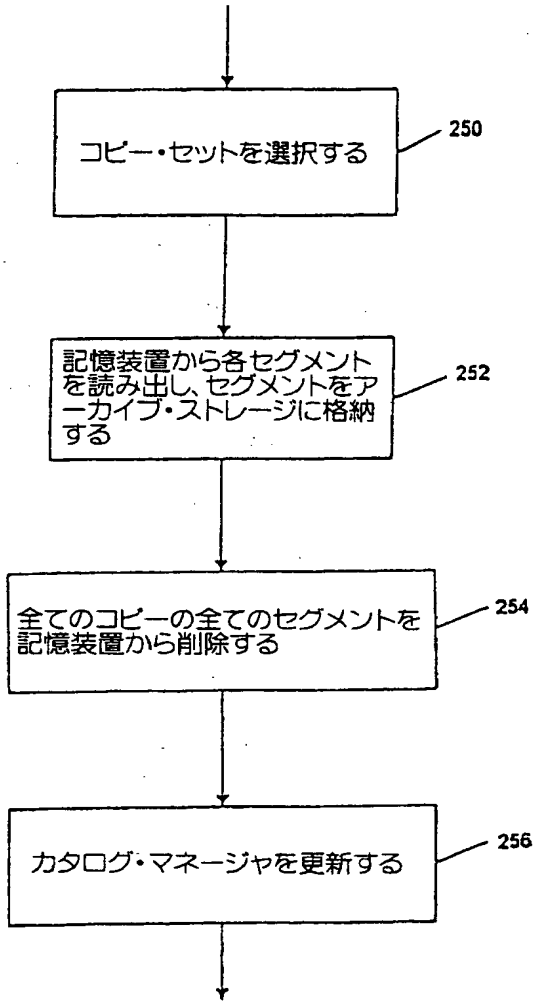
【FIG. 7】

【図８】


開始
次のファイルを選択する
230
選択したファイルの記憶装置上にある全セグメントを識別する
232
次のセグメントを選択する
234
新たな記憶装置を割り当てる
235
セグメントを読み出し、新たな記憶装置にセグメントを書き込む
236
古い記憶装置からセグメントを削除し、セグメント・テーブルを更新する
237
全セグメントを処理したか？
238
No
Yes
セグメント・テーブルをカタログ・マネージャに送る
239
全ファイルを処理したか？
240
No
Yes
終了

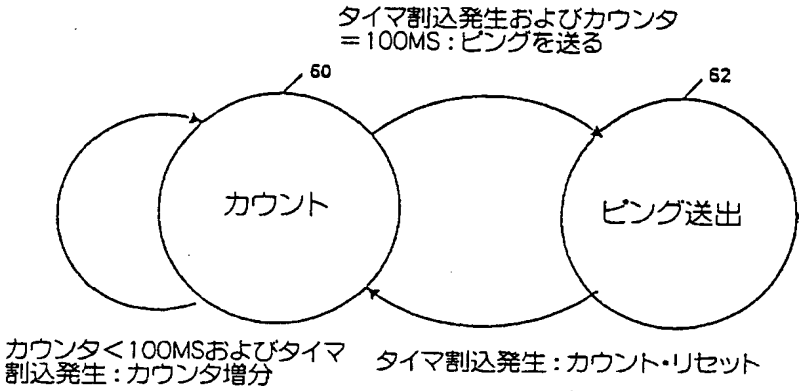
【FIG. 8】

【図９】


コピー・セットを選択する
250
記憶装置から各セグメントを読み出し、セグメントをアーカイブ・ストレージに格納する
252
全てのコピーの全てのセグメントを記憶装置から削除する
254
カタログ・マネージャを更新する
256

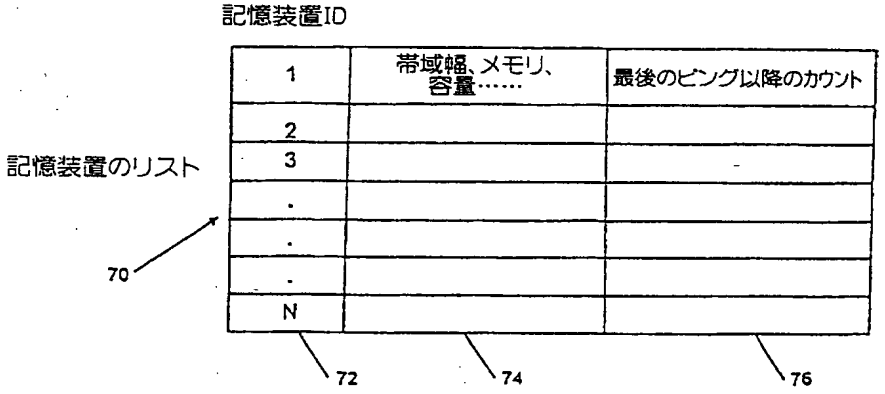
【FIG. 9】

【図１０】



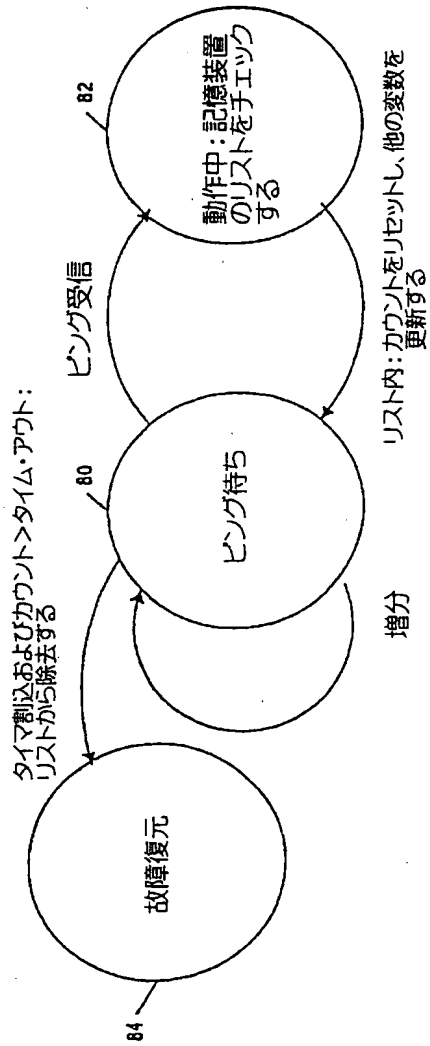
【FIG. 10】

【図１１】



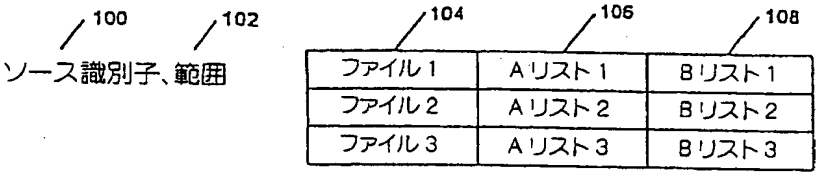
【FIG. 11】

【図１２】



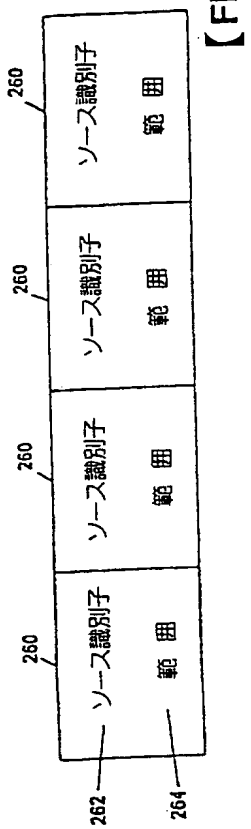
【FIG. 12】

【図１３】



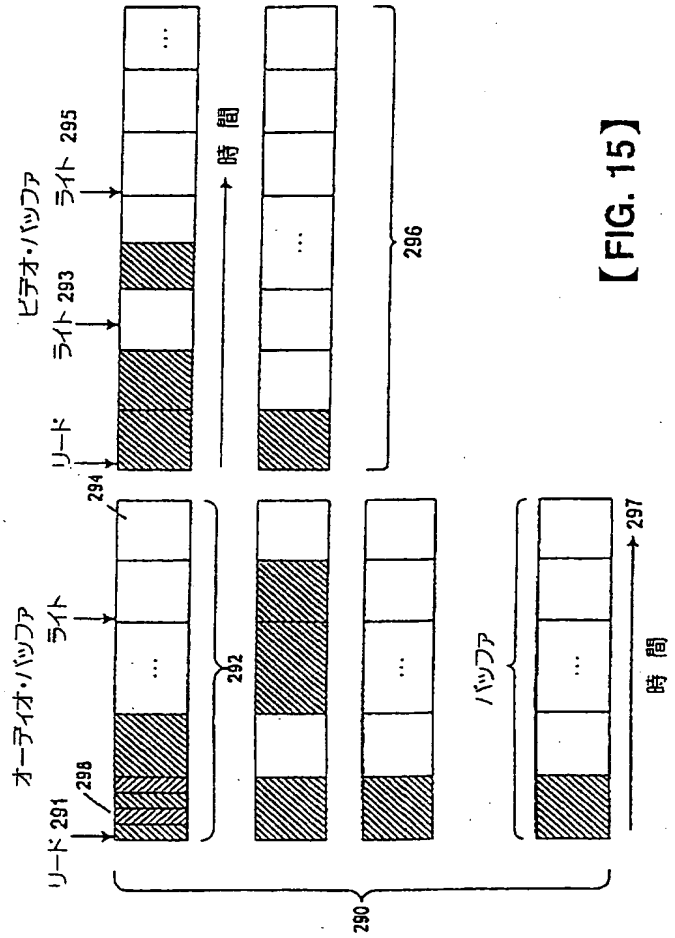
【FIG. 13】

【図１４】


260
ソース識別子
範 囲
260
ソース識別子
範 囲
260
ソース識別子
範 囲
260
ソース識別子
範 囲
262
264

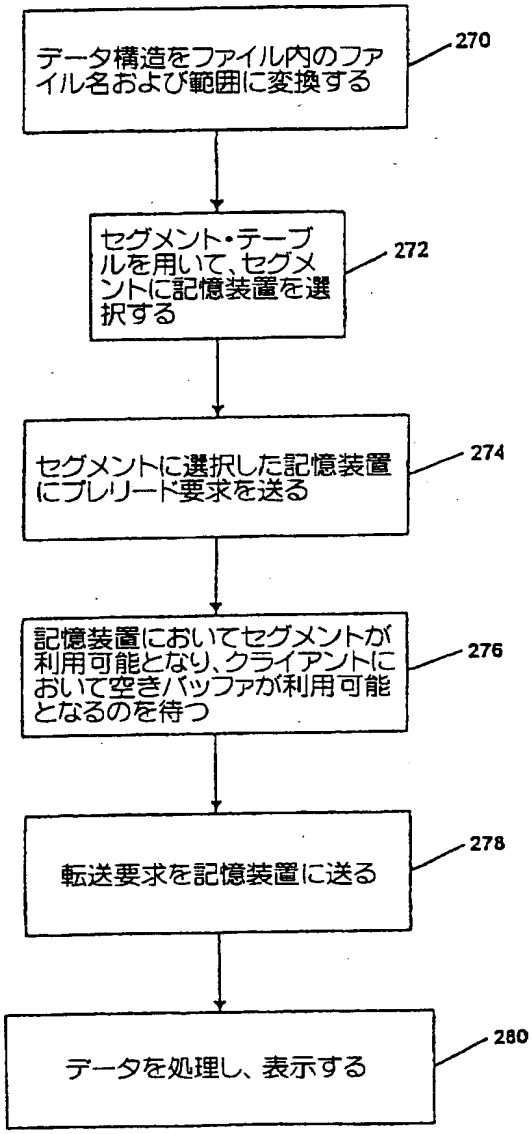
【FIG. 14】

【図１５】


ビデオ・バッファ
ライト 295
ライト 293
リード
294
時間
296
オーディオ・バッファ
ライト
298
リード 291
292
バッファ
297
時間
290

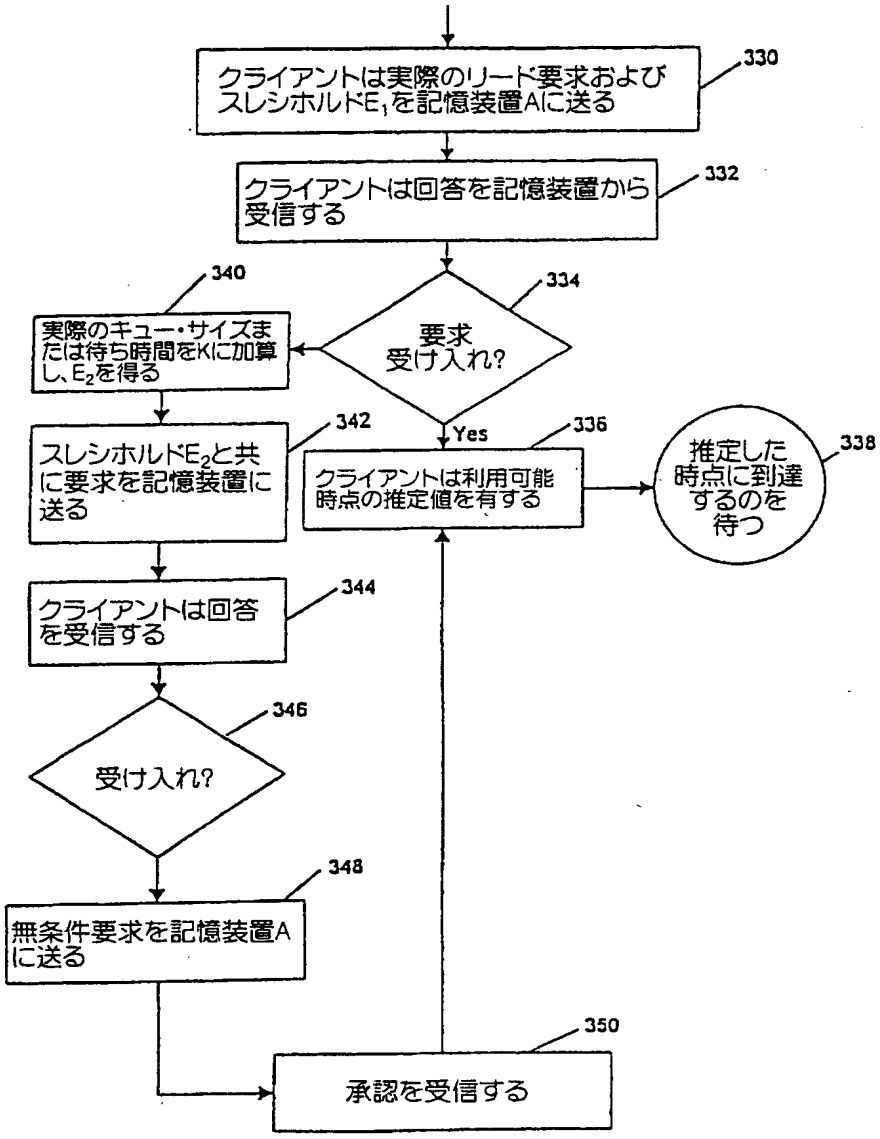
【FIG. 15】

【図１６】


データ構造をファイル内のファイル名および範囲に変換する
270
セグメント・テーブルを用いて、セグメントに記憶装置を選択する
272
セグメントに選択した記憶装置にプレリード要求を送る
274
記憶装置においてセグメントが利用可能となり、クライアントにおいて空きバッファが利用可能となるのを待つ
276
転送要求を記憶装置に送る
278
データを処理し、表示する
280

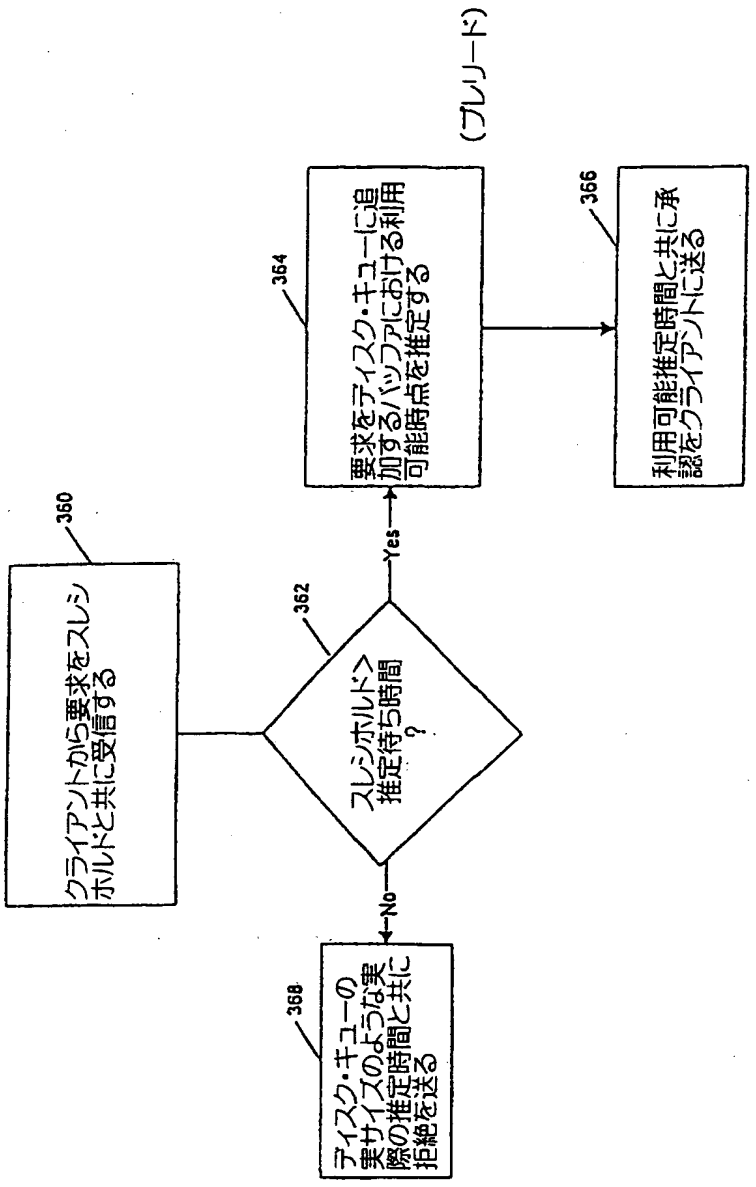
【FIG. 16】

【図１７】


クライアントは実際のリード要求およびスレシホルドE1を記憶装置Aに送る
330
クライアントは回答を記憶装置から受信する
332
要求受け入れ?
334
Yes
実際のキュー・サイズまたは待ち時間をKに加算し、E2を得る
340
スレシホルドE2と共に要求を記憶装置に送る
342
クライアントは利用可能時点の推定値を有する
336
推定した時点に到達するのを待つ
338
クライアントは回答を受信する
344
受け入れ?
346
無条件要求を記憶装置Aに送る
348
承認を受信する
350

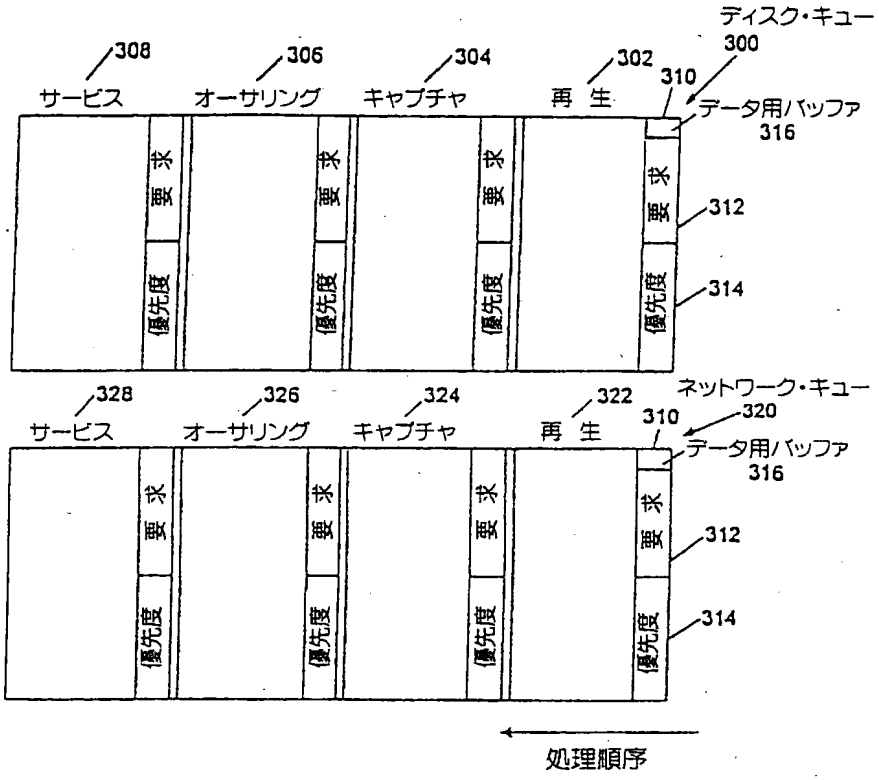
【FIG. 17】

【図１８】


(プレリード)
360
クライアントから要求をスレシホルドと共に受信する
362
スレシホルド>推定待ち時間?
Yes
No
364
要求をディスク・キューに追加する/バッファにおける利用可能時点を推定する
366
利用可能推定時間と共に承認をクライアントに送る
368
ディスク・キューの実サイズのような実際の推定時間と共に拒絶を送る

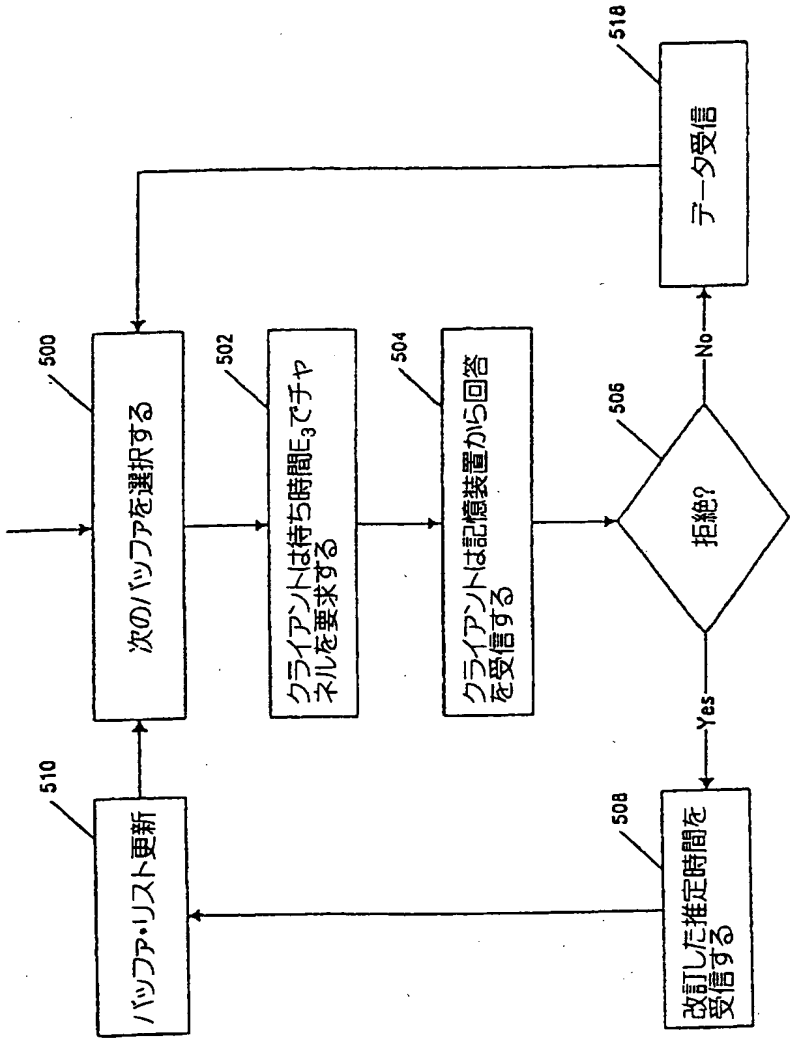
【FIG. 18】

【図１９】


ディスク・キュー
300
308
306
304
302
310
サービス
オーサリング
キャプチャ
再 生
データ用バッファ
316
要求
優先度
312
314
ネットワーク・キュー
328
326
324
322
320
処理順序

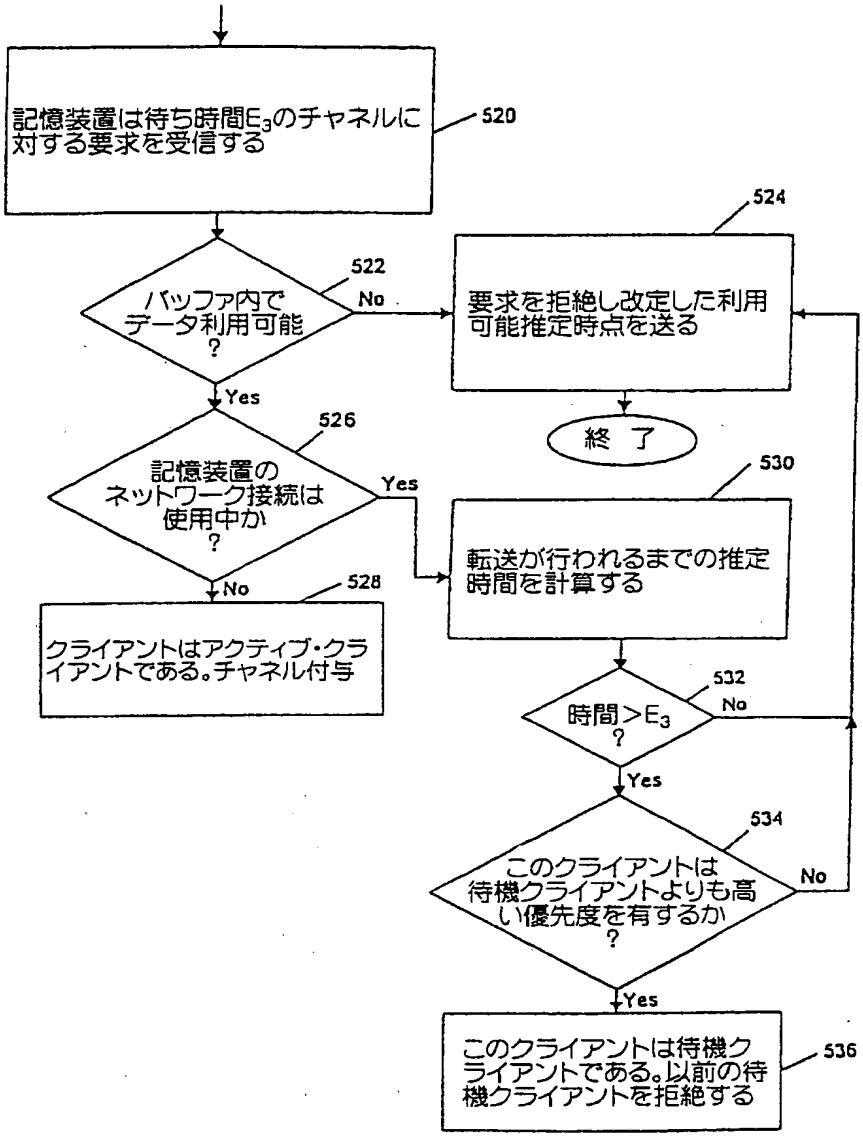
【FIG. 19】

【図２０】


500
次のバッファを選択する
502
クライアントは待ち時間E_3でチャネルを要求する
504
クライアントは記憶装置から回答を受信する
506
拒絶?
No
Yes
508
改訂した推定時間を受信する
510
バッファ・リスト更新
518
データ受信

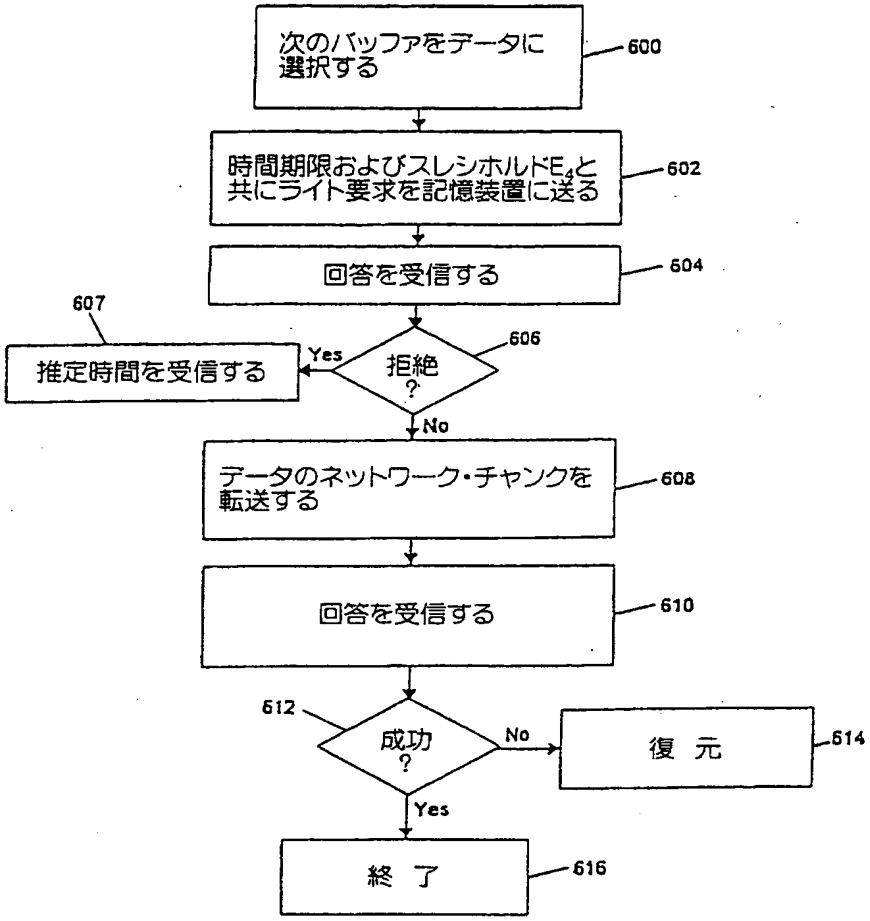
【FIG. 20】

【図２１】


記憶装置は待ち時間E_3のチャネルに対する要求を受信する
520
524
バッファ内でデータ利用可能？
522
No
要求を拒絶し改定した利用可能推定時点を送る
Yes
終了
526
記憶装置のネットワーク接続は使用中か？
Yes
530
転送が行われるまでの推定時間を計算する
No
528
クライアントはアクティブ・クライアントである。チャネル付与
時間＞E_3？
532
No
Yes
534
このクライアントは待機クライアントよりも高い優先度を有するか？
No
Yes
このクライアントは待機クライアントである。以前の待機クライアントを拒絶する
536

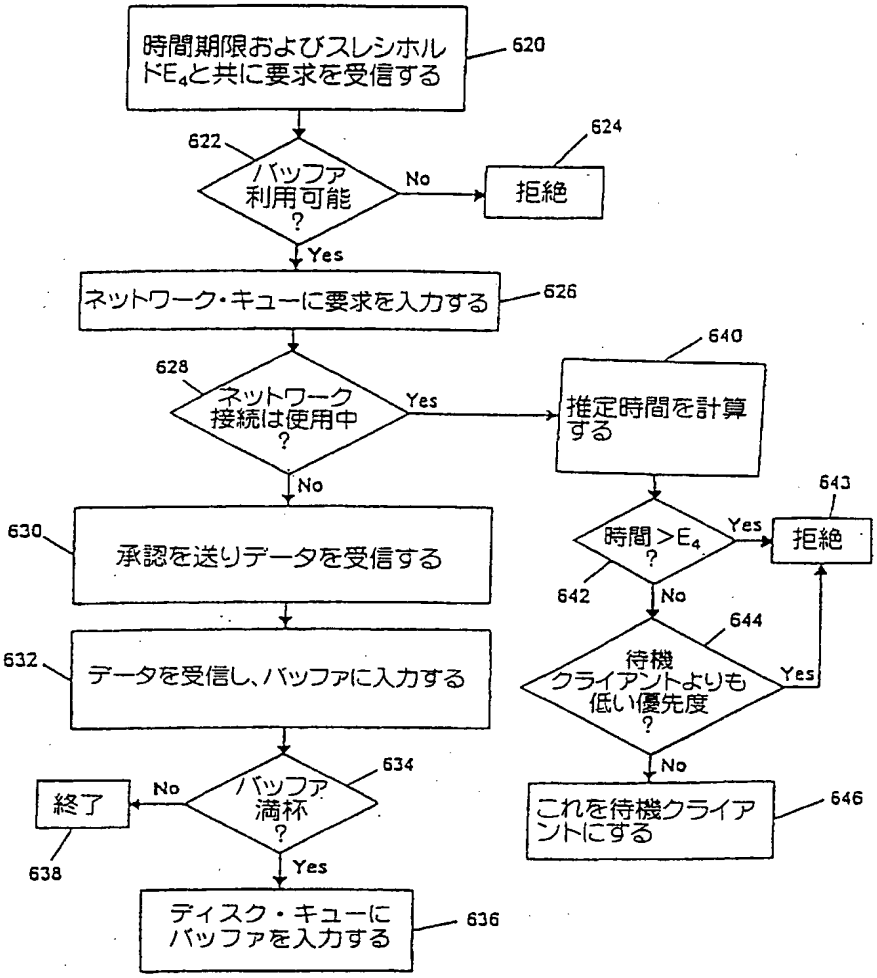
【FIG. 21】

【図２２】


次のバッファをデータに選択する
600
時間期限およびスレシホルドE4と共にライト要求を記憶装置に送る
602
回答を受信する
604
607
推定時間を受信する
Yes
拒絶?
606
No
データのネットワーク・チャンクを転送する
608
回答を受信する
610
612
成功?
No
復元
614
Yes
終了
616

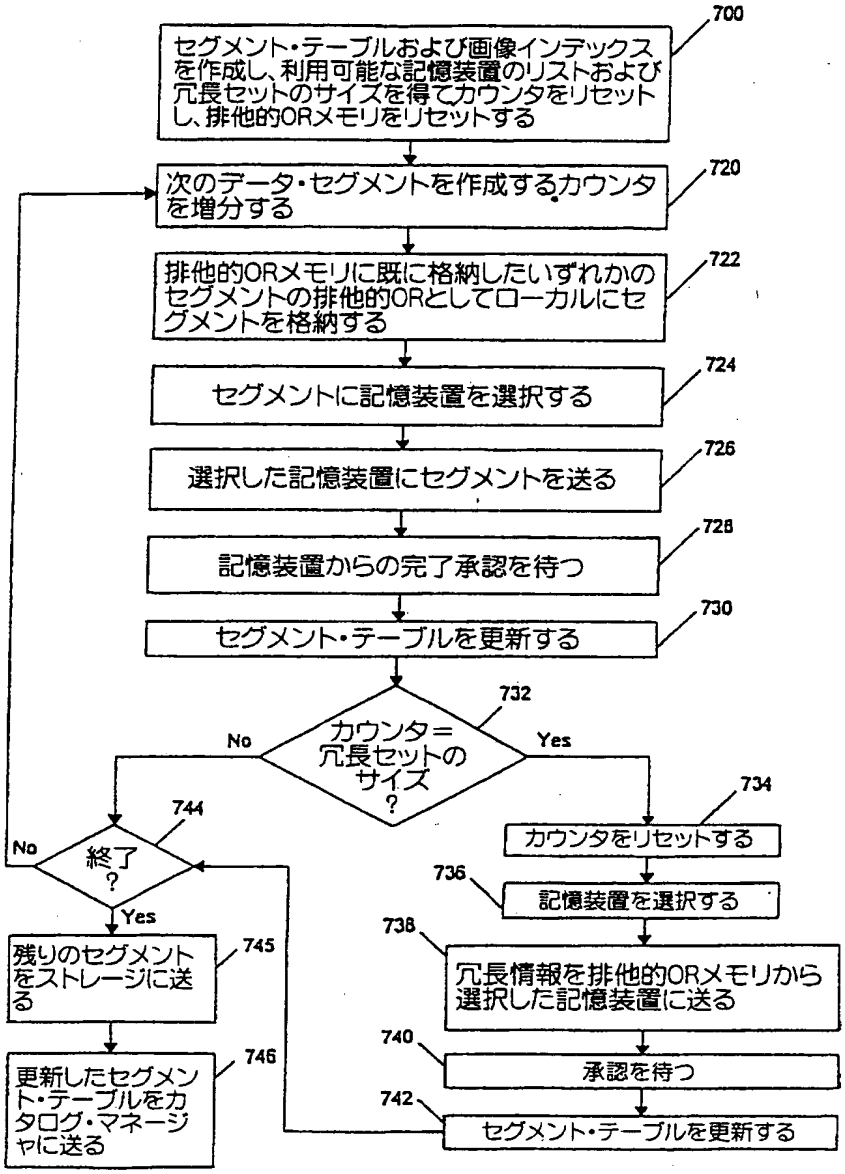
【FIG. 22】

【図２３】


時間期限およびスレシホルドE4と共に要求を受信する
620
622
バッファ利用可能？
No
624
拒絶
Yes
ネットワーク・キューに要求を入力する
626
628
ネットワーク接続は使用中？
Yes
640
推定時間を計算する
No
643
時間>E4？
Yes
拒絶
642
No
644
待機クライアントよりも低い優先度？
Yes
No
646
これを待機クライアントにする
630
承認を送りデータを受信する
632
データを受信し、バッファに入力する
634
バッファ満杯？
No
終了
638
Yes
ディスク・キューにバッファを入力する
636

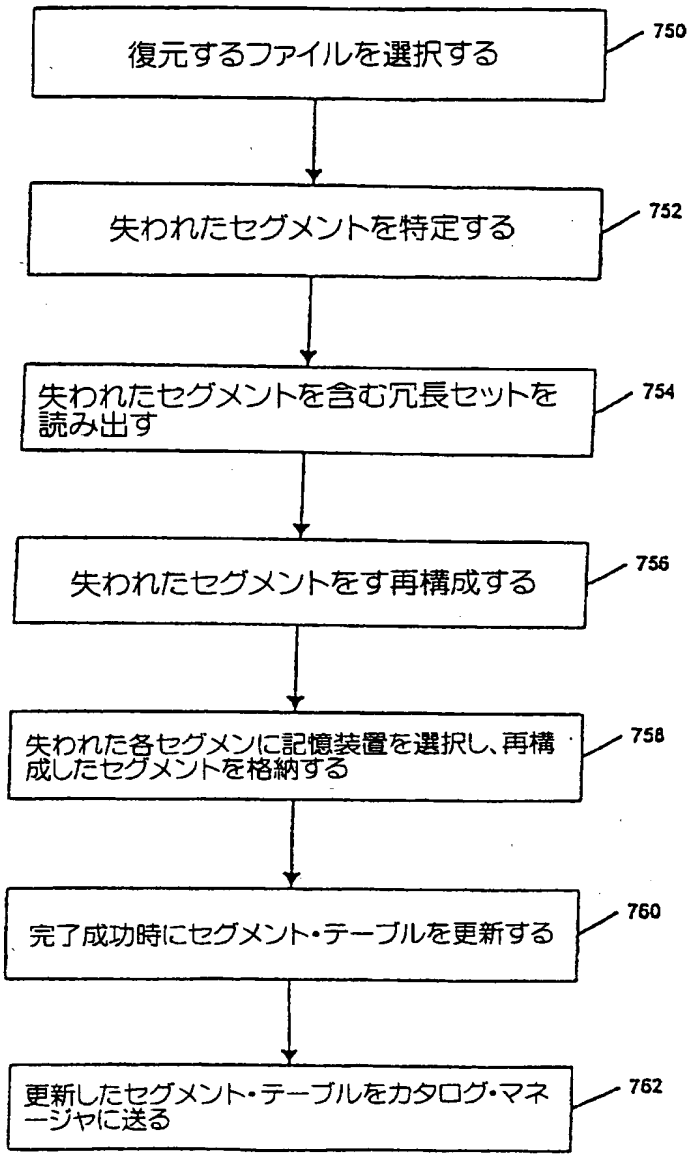
【FIG. 23】

【図２４】


700
セグメント・テーブルおよび画像インデックスを作成し、利用可能な記憶装置のリストおよび冗長セットのサイズを得てカウンタをリセットし、排他的ORメモリをリセットする
720
次のデータ・セグメントを作成するカウンタを増分する
722
排他的ORメモリに既に格納したいずれかのセグメントの排他的ORとしてローカルにセグメントを格納する
724
セグメントに記憶装置を選択する
726
選択した記憶装置にセグメントを送る
728
記憶装置からの完了承認を待つ
730
セグメント・テーブルを更新する
732
カウンタ＝冗長セットのサイズ？
No
Yes
734
カウンタをリセットする
736
記憶装置を選択する
738
冗長情報を排他的ORメモリから選択した記憶装置に送る
740
承認を待つ
742
セグメント・テーブルを更新する
744
終了？
No
Yes
745
残りのセグメントをストレージに送る
746
更新したセグメント・テーブルをカタログ・マネージャに送る

【FIG. 24】

【図２５】


復元するファイルを選択する
750
失われたセグメントを特定する
752
失われたセグメントを含む冗長セットを読み出す
754
失われたセグメントをす再構成する
756
失われた各セグメンに記憶装置を選択し、再構成したセグメントを格納する
758
完了成功時にセグメント・テーブルを更新する
760
更新したセグメント・テーブルをカタログ・マネージャに送る
762

【FIG. 25】

【手続補正書】特許協力条約第３４条補正の翻訳文提出書

【提出日】平成１２年４月７日（２０００．４．７）

【手続補正１】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項１】　コンピュータ用ファイル・システムであって、前記コンピュータが、当該コンピュータ上で実行するアプリケーションからの要求に応答してコンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上に格納してあるデータを読み出すことを可能とし、ファイルは、データ・セグメントと各セグメント毎に対応する冗長情報とを含み、各ファイル毎に、各データ・セグメントを前記記憶装置のランダムまたは疑似ランダムに選択した１つに格納し、各データ・セグメント毎に、前記記憶装置のランダムまたは疑似ランダムに選択した１つに前記対応する冗長情報を格納し、セグメントに対する前記冗長情報は、前記セグメントの少なくとも１つのコピーであり、

前記要求に応答してデータを読み出し、前記選択したデータの各セグメント毎に、前記セグメントを格納する記憶装置の１つを選択する手段と、

前記要求したデータの各セグメントを、当該セグメントに対して選択した記憶装置から、読み出す手段と、

各セグメント毎に、前記選択した記憶装置から読み出した選択データのセグメントを直列化する手段と、

前記特定した記憶装置からデータを受信したとき、前記アプリケーションに前記データを供給する手段と、

を備えるファイル・システム。

【請求項２】　前記記憶装置から１つを選択する前記手段は、前記複数の記憶装置上の要求の負荷が実質的に均衡化するように、前記記憶装置を選択する請求項１記載のファイル・システム。

【請求項３】　前記選択手段は、前記セグメントのためのどの記憶装置が、前記要求に応じるための推定時間が最も短いかについての推定に基づいて、前記セグメントのための前記記憶装置を選択する請求項２記載のファイル・システム。

【請求項４】　前記選択手段が、

前記ファイル・システムにおいて、

前記記憶装置の１つからデータを要求し、推定時間を示す手段と、

前記第１記憶装置が前記要求を拒絶した場合、前記記憶装置の他のものからデータを要求し、推定時間を示す手段と、

前記第２記憶装置が前記要求を拒絶した場合、前記第１記憶装置から前記データを要求する手段と、

を含み、

各記憶装置において、

前記要求が前記推定時間以内に前記記憶装置によって対応することができない場合、データの要求を拒絶する手段と、

前記要求が前記推定時間以内に前記記憶装置によって対応することができる場合、データの要求を受け入れる手段と、

を含む請求項３記載のファイル・システム。

【請求項５】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項２記載のファイル・システム。

【請求項６】　転送をスケジュールする前記手段は、

前記ファイル・システムにおいて、

前記選択した記憶装置から前記データの転送を要求し、待ち時間を示す手段と、

前記選択した記憶装置が前記データを転送する前記要求を拒絶した場合、別の記憶装置から前記データを要求する手段と、

を含み、

前記記憶装置において、

前記データが前記指示した待ち時間までに前記記憶装置から転送することができない場合、データ転送要求を拒絶する手段と、

前記選択した記憶装置が前記待ち時間以内に前記データを転送することができる場合、前記データを転送する手段と、

を含む請求項５記載のファイル・システム。

【請求項７】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項１記載のファイル・システム。

【請求項８】　各セグメントを読み出す前記手段は、前記選択した記憶装置からの前記データの転送をスケジュールし、前記記憶装置が効率的にデータ転送を行なうようにする手段を備える請求項１記載のファイル・システム。

【請求項９】　転送をスケジュールする前記手段は、

前記ファイル・システムにおいて、

前記選択した記憶装置から前記データの転送を要求し、待ち時間を示す手段と、

前記選択した記憶装置が前記データを転送する前記要求を拒絶した場合、別の記憶装置から前記データを要求する手段と

を含み、

前記記憶装置において、

前記データが前記指示した待ち時間までに前記記憶装置から転送することができない場合、データ転送要求を拒絶する手段と、

前記選択した記憶装置が前記待ち時間以内に前記データを転送することができる場合、前記データを転送する手段と、

を含む請求項８記載のファイル・システム。

【請求項１０】　各セグメントを前記記憶装置の異なるものに格納する請求項１記載のファイル・システム。

【請求項１１】　各セグメントの各コピーを、前記記憶装置の相対的仕様の関数として定義した確率分布にしたがって、前記複数の記憶装置の１つに割り当てる請求項１０記載のファイル・システム。

【請求項１２】　更に、コンピュータ読み取り可能ロジックを格納し、データ・セグメントの指示を用いてコンピュータによるアクセスが可能なセグメント・テーブルを定義するコンピュータ読み取り可能媒体を備え、前記セグメントおよび対応する冗長情報を格納した前記複数の記憶装置から、該記憶装置の指示を検索する請求項１記載の分散データ記憶システム。

【請求項１３】　前記複数の記憶装置が、

前記コンピュータ・ネットワークに接続した第１記憶装置と、

前記コンピュータ・ネットワークに接続した第２記憶装置と、

前記コンピュータ・ネットワークに接続した第３記憶装置と、

を備える請求項１記載のファイル・システム。

【請求項１４】　コンピュータ用ファイル・システムであって、前記コンピュータが、当該コンピュータ上で実行するアプリケーションからの要求に応答してコンピュータ・ネットワークを通じて独立したリモート記憶装置にアクセスし、前記記憶装置上にデータを格納することを可能とするファイル・システムにおいて、

前記データを格納する前記要求に応答して、前記データを複数のセグメントに分割する手段と、

各セグメント毎に記憶装置をランダムまたは疑似ランダムに選択し、前記選択した記憶装置に前記データを格納するように要求する手段と、

各セグメント毎に、当該セグメントに対応する冗長情報のための記憶装置をランダムまたは疑似ランダムに選択し、前記選択した記憶装置に前記冗長情報を格納するように要求する手段であって、前記冗長情報が前記セグメントの少なくとも１つのコピーである、手段と、

前記データを格納したか否かについて前記アプリケーションに確認する手段と、

を備えるファイル・システム。

【請求項１５】　記選択する手段は、前記記憶装置の部分集合を選択する手段と、前記選択した部分集合内にある前記記憶装置の中から少なくとも２つの前記記憶装置を選択する手段とを含む請求項１４記載のファイル・システム。

【請求項１６】　各セグメントの各コピーは、前記記憶装置の相対的仕様の関数として定義した、確率分布にしたがって、前記複数の記憶装置の１つに割り当てられる請求項１４記載のファイル・システム。

【請求項１７】　更に、コンピュータ読み取り可能ロジックを格納し、データ・セグメントの指示を用いてコンピュータによるアクセスが可能なセグメント・テーブルを定義するコンピュータ読み取り可能媒体を備え、前記セグメントおよび対応する冗長情報を格納した前記複数の記憶装置から、該記憶装置の指示を検索する請求項１４記載のファイル・システム。

【請求項１８】　前記複数の記憶装置が、

前記コンピュータ・ネットワークに接続した第１記憶装置と、

前記コンピュータ・ネットワークに接続した第２記憶装置と、

前記コンピュータ・ネットワークに接続した第３記憶装置と、

を備える請求項１４記載のファイル・システム。

【手続補正書】

【提出日】平成１３年３月５日（２００１．３．５）

【手続補正１】

【補正対象書類名】図面

【補正対象項目名】図１

【補正方法】変更

【補正内容】

【図１】

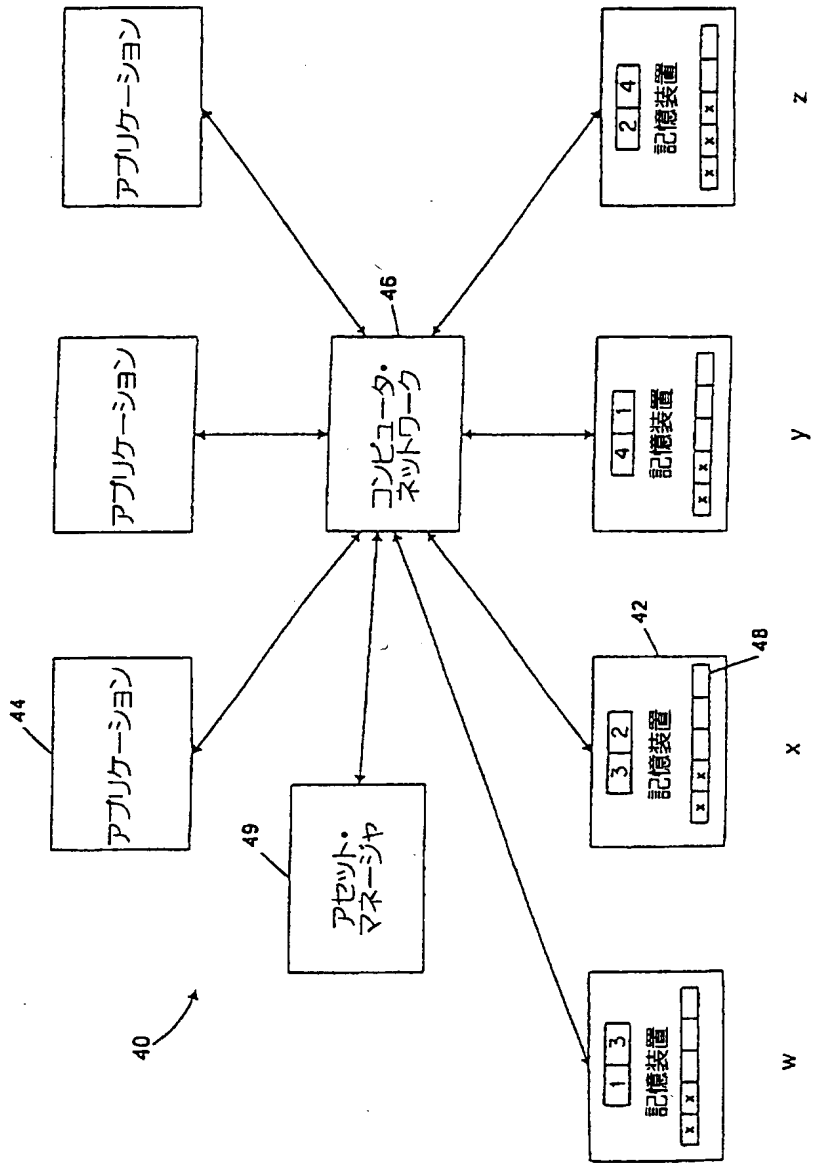

アプリケーション
アプリケーション
アプリケーション
コンピュータ・ネットワーク
アセット・マネージャ
記憶装置
記憶装置
記憶装置
記憶装置
40
42
44
46
48
49
w
x
y
z

FIG. 1A

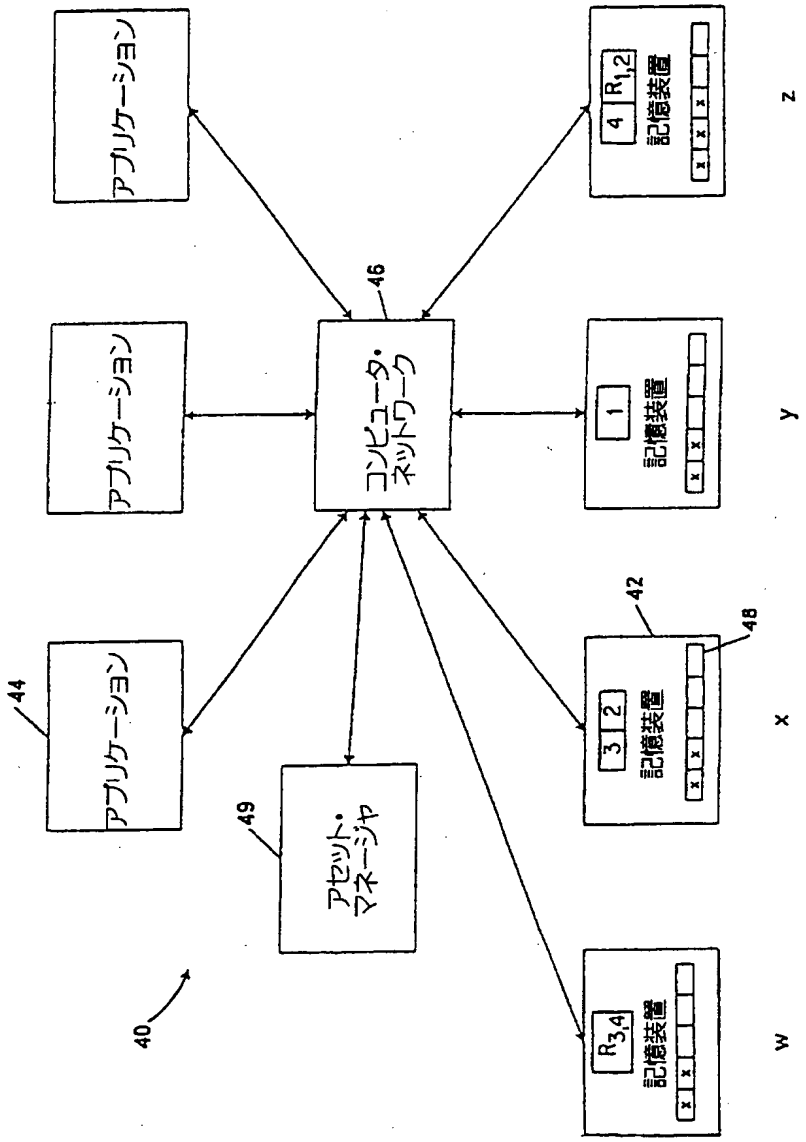


FIG. 1B

【手続補正２】

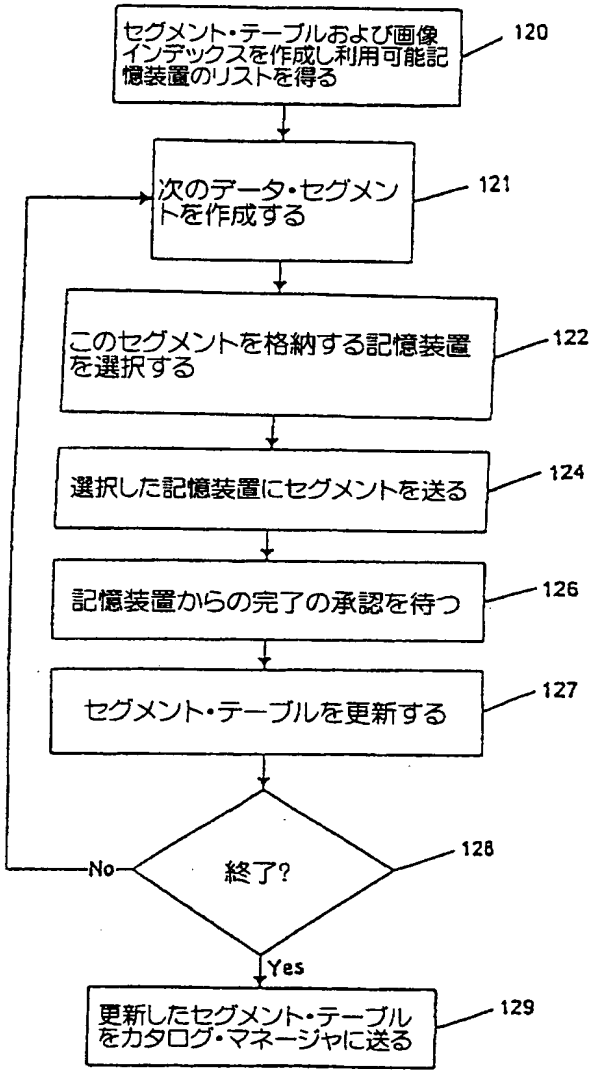
【補正対象書類名】図面

【補正対象項目名】図３

【補正方法】変更

【補正内容】

【図３】



【手続補正３】

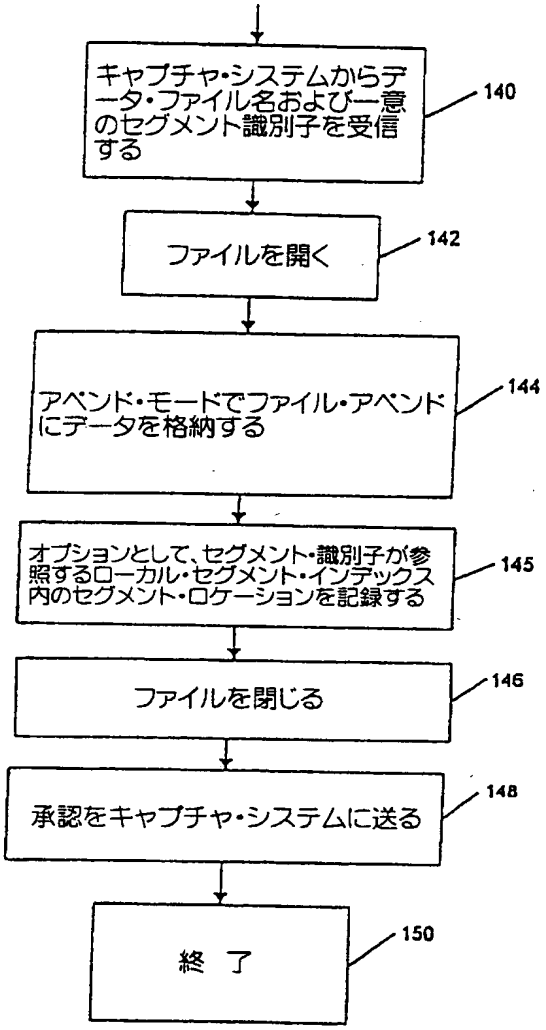
【補正対象書類名】図面

【補正対象項目名】図４

【補正方法】変更

【補正内容】

【図４】



【手続補正４】

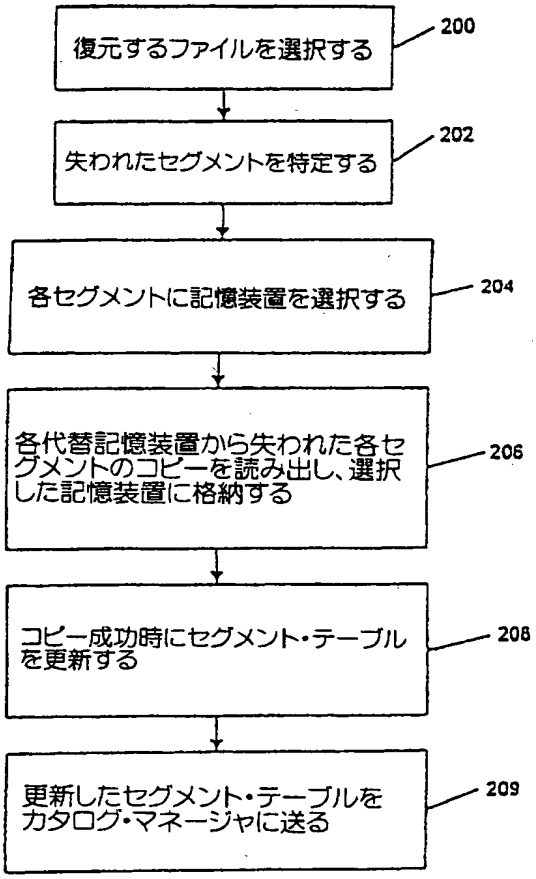
【補正対象書類名】図面

【補正対象項目名】図５

【補正方法】変更

【補正内容】

【図５】



【手続補正５】

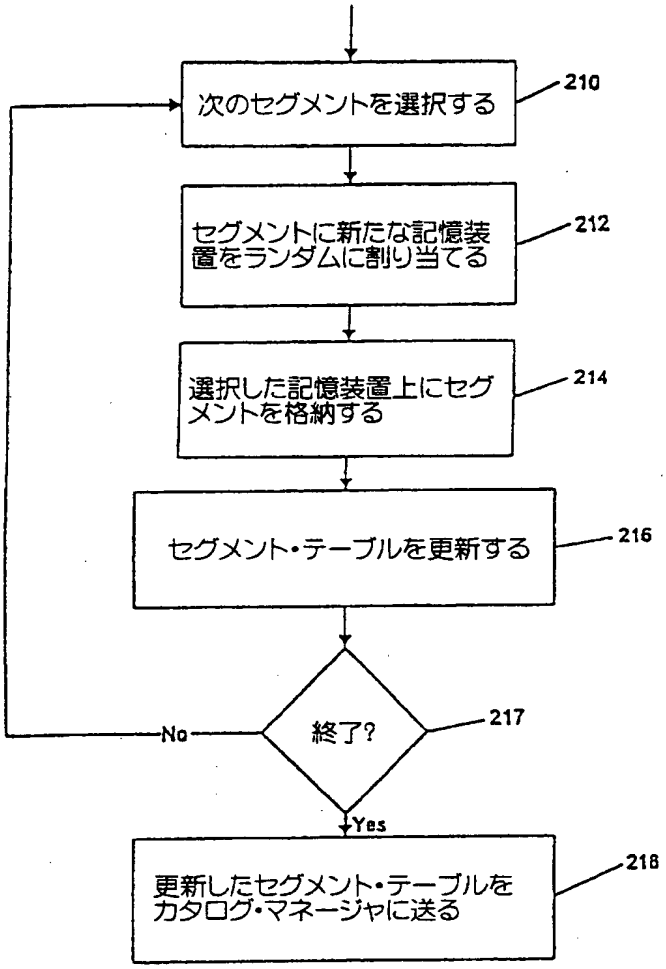
【補正対象書類名】図面

【補正対象項目名】図６

【補正方法】変更

【補正内容】

【図６】



【手続補正６】

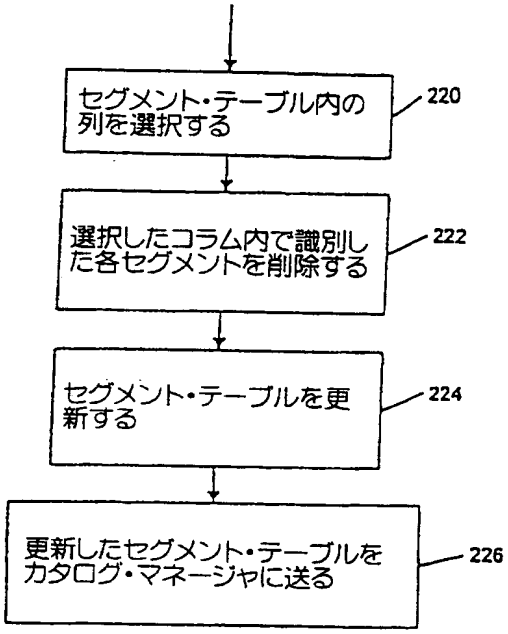
【補正対象書類名】図面

【補正対象項目名】図７

【補正方法】変更

【補正内容】

【図７】



【手続補正７】

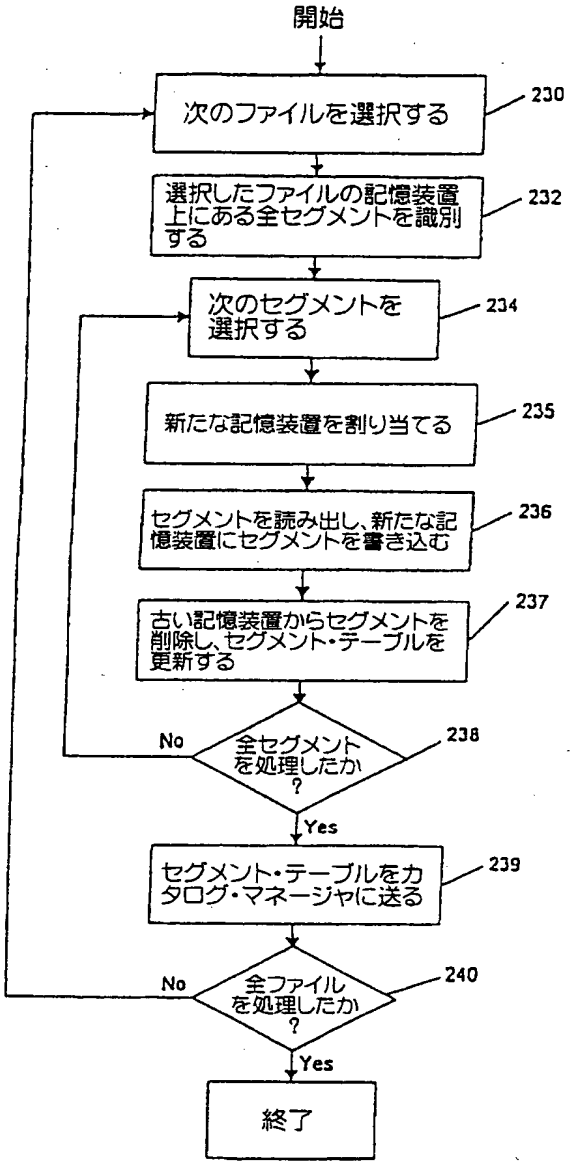
【補正対象書類名】図面

【補正対象項目名】図８

【補正方法】変更

【補正内容】

【図８】



【手続補正８】

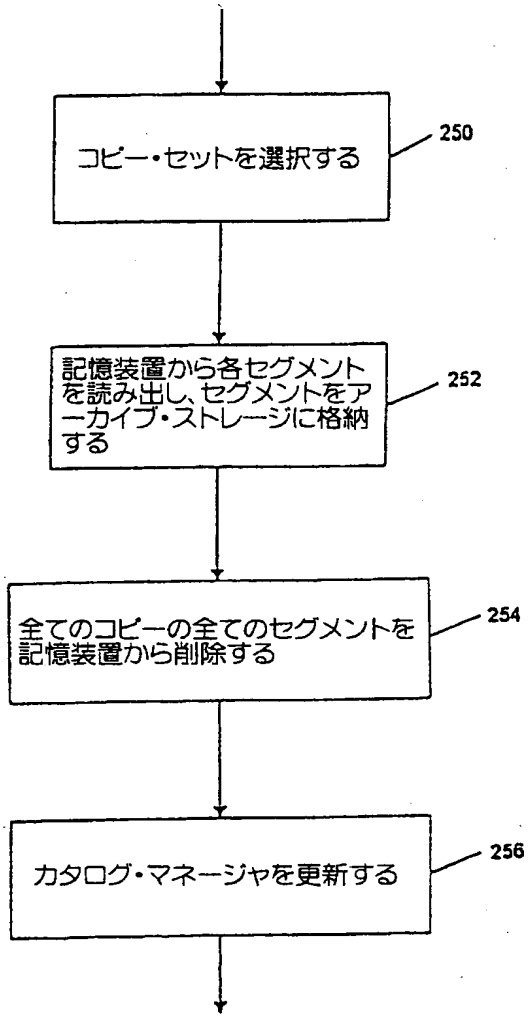
【補正対象書類名】図面

【補正対象項目名】図９

【補正方法】変更

【補正内容】

【図９】



【手続補正９】

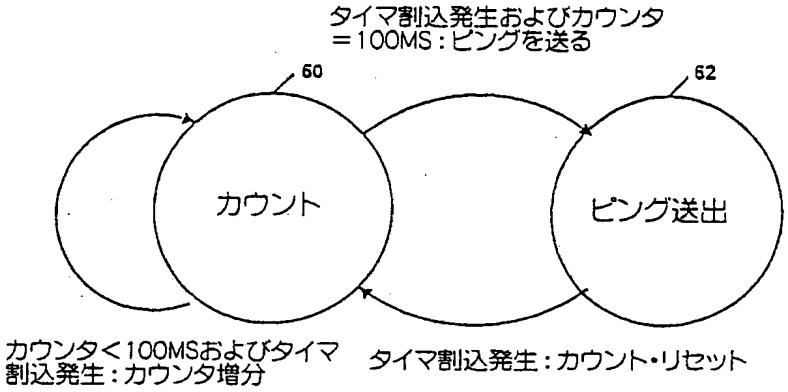
【補正対象書類名】図面

【補正対象項目名】図１０

【補正方法】変更

【補正内容】

【図１０】



【手続補正１０】

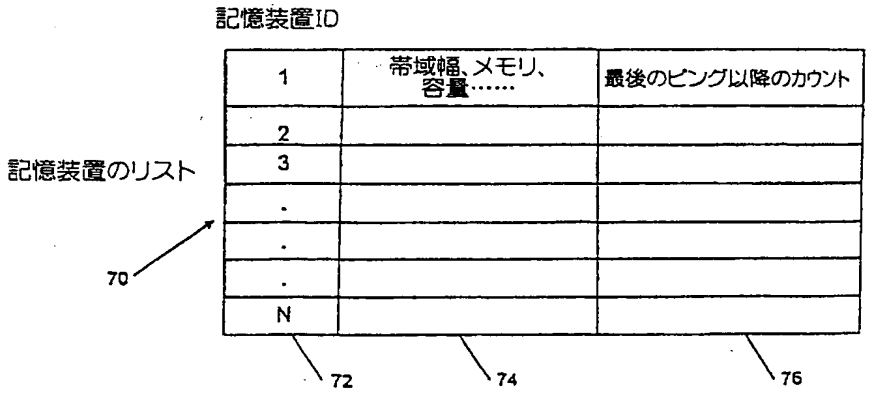
【補正対象書類名】図面

【補正対象項目名】図１１

【補正方法】変更

【補正内容】

【図１１】



【手続補正１１】

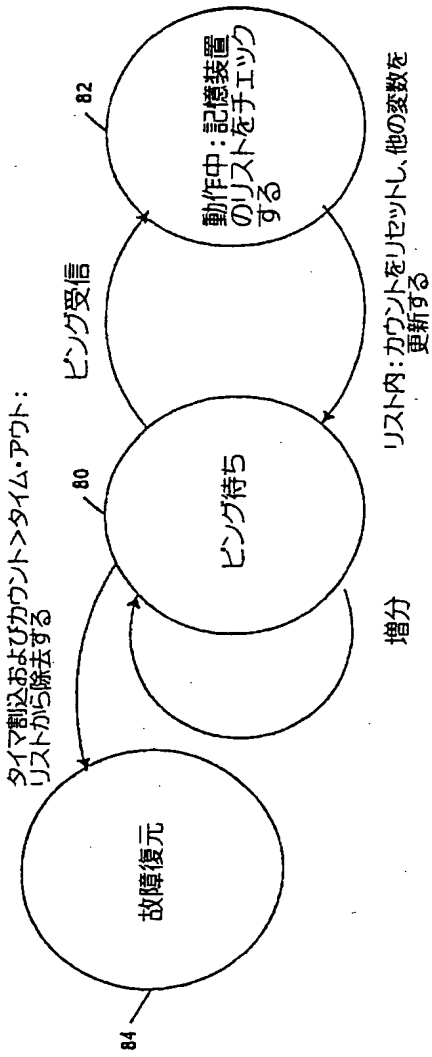
【補正対象書類名】図面

【補正対象項目名】図１２

【補正方法】変更

【補正内容】

【図１２】



【手続補正１２】

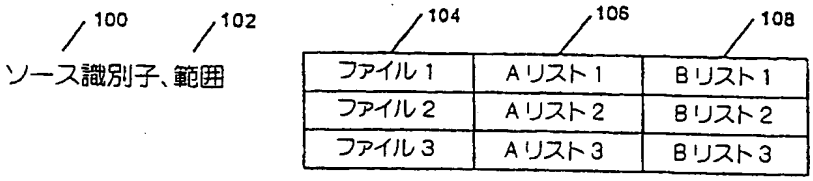
【補正対象書類名】図面

【補正対象項目名】図１３

【補正方法】変更

【補正内容】

【図１３】



【手続補正１３】

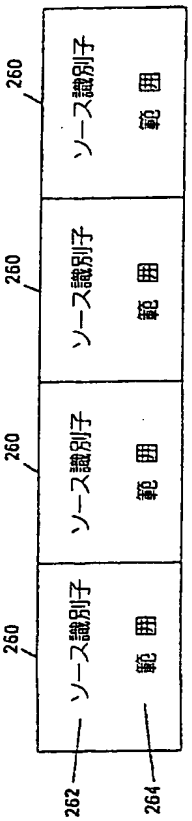
【補正対象書類名】図面

【補正対象項目名】図１４

【補正方法】変更

【補正内容】

【図１４】



【手続補正１４】

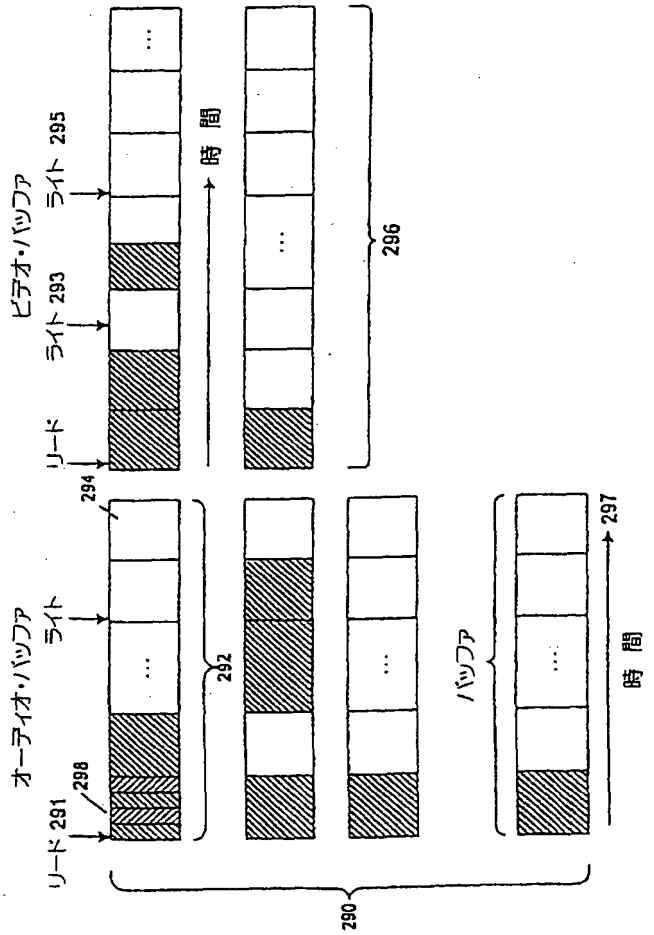
【補正対象書類名】図面

【補正対象項目名】図１５

【補正方法】変更

【補正内容】

【図１５】



【手続補正１５】

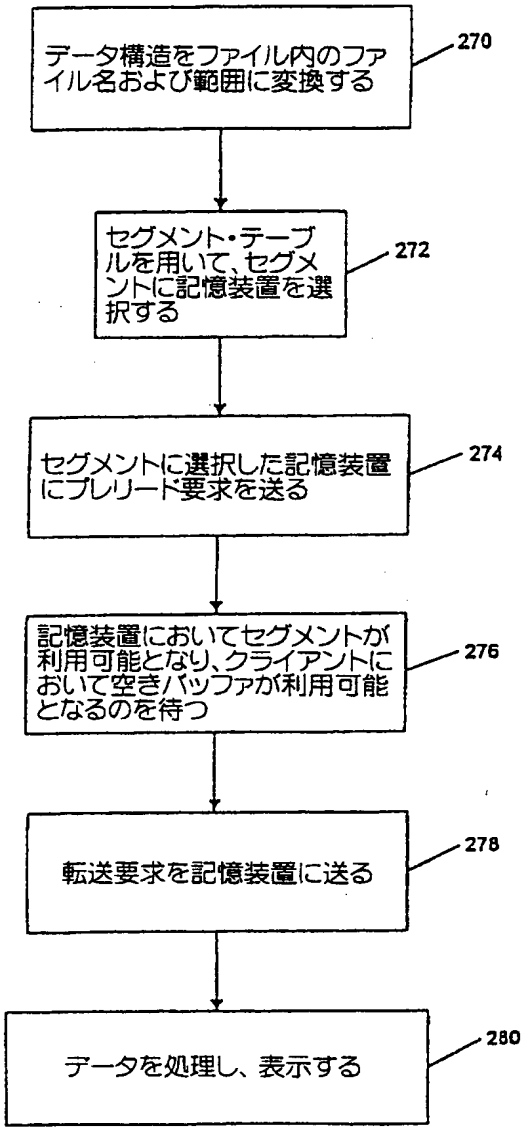
【補正対象書類名】図面

【補正対象項目名】図１６

【補正方法】変更

【補正内容】

【図１６】



【手続補正１６】

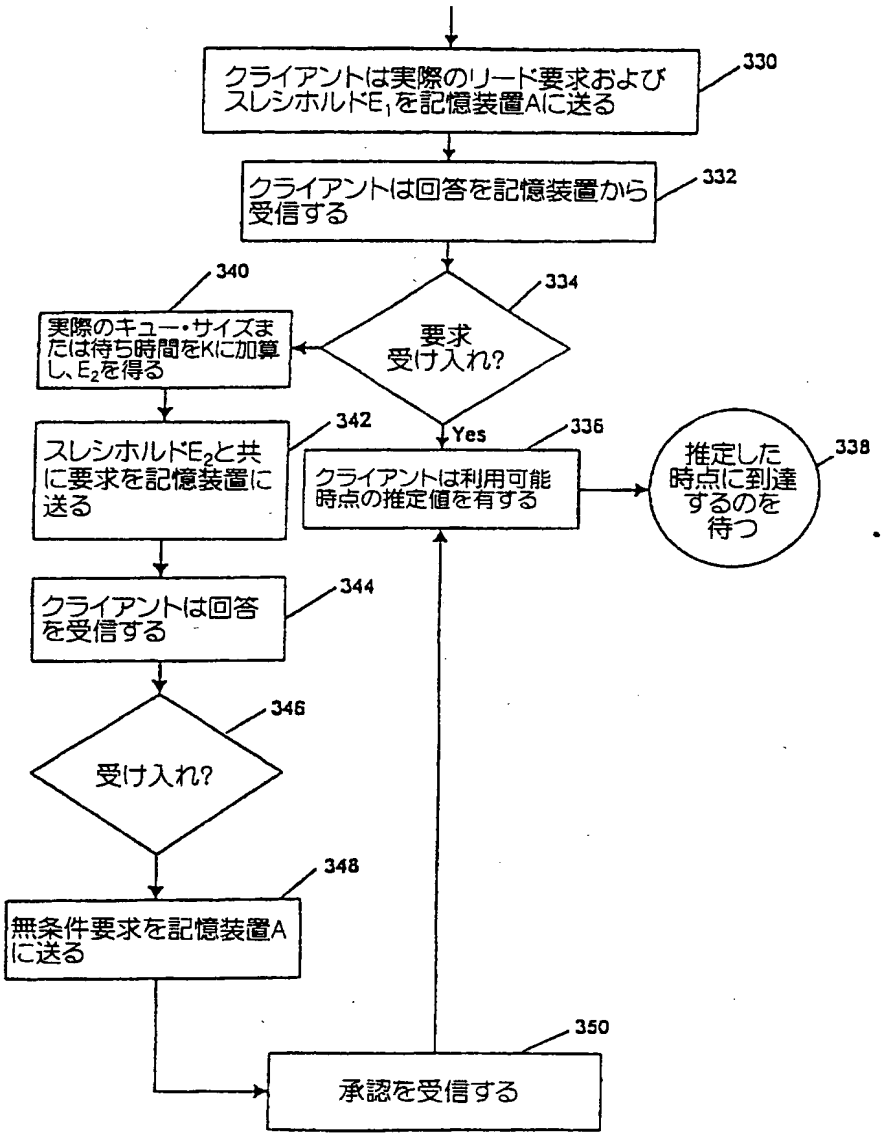
【補正対象書類名】図面

【補正対象項目名】図１７

【補正方法】変更

【補正内容】

【図１７】



【手続補正１７】

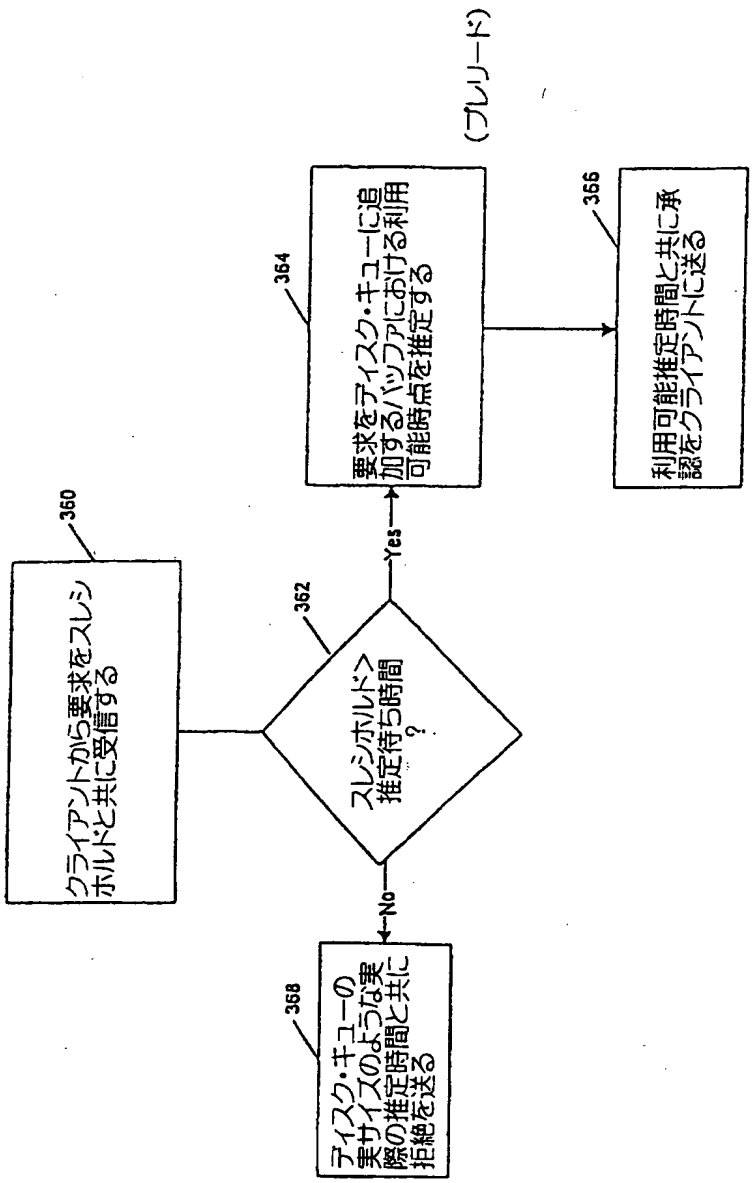
【補正対象書類名】図面

【補正対象項目名】図１８

【補正方法】変更

【補正内容】

【図１８】



【手続補正１８】

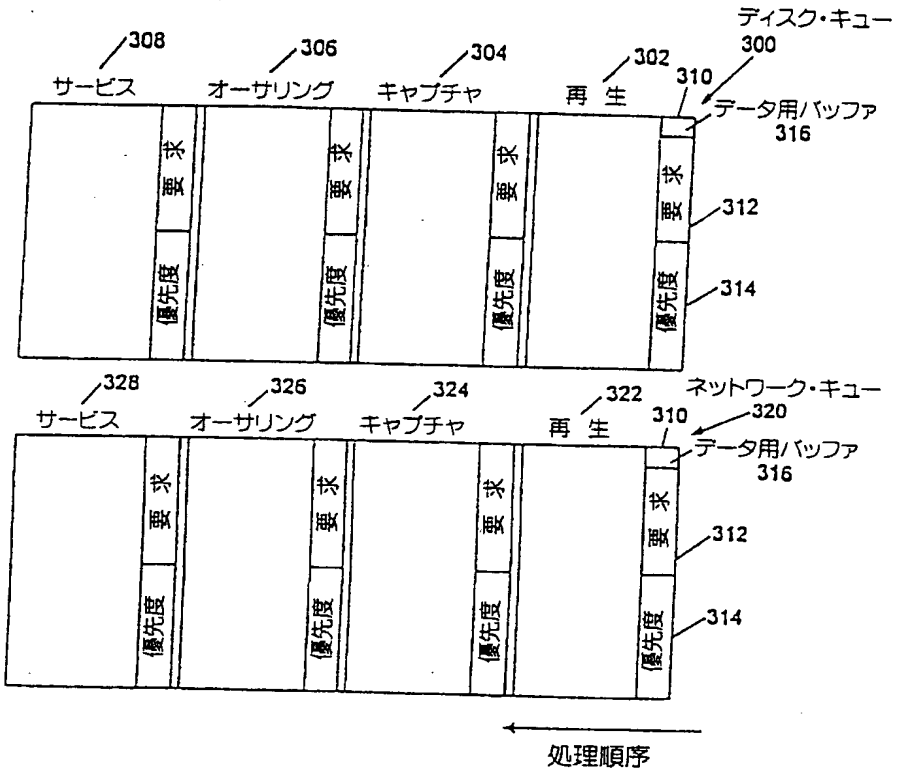
【補正対象書類名】図面

【補正対象項目名】図１９

【補正方法】変更

【補正内容】

【図１９】



【手続補正１９】

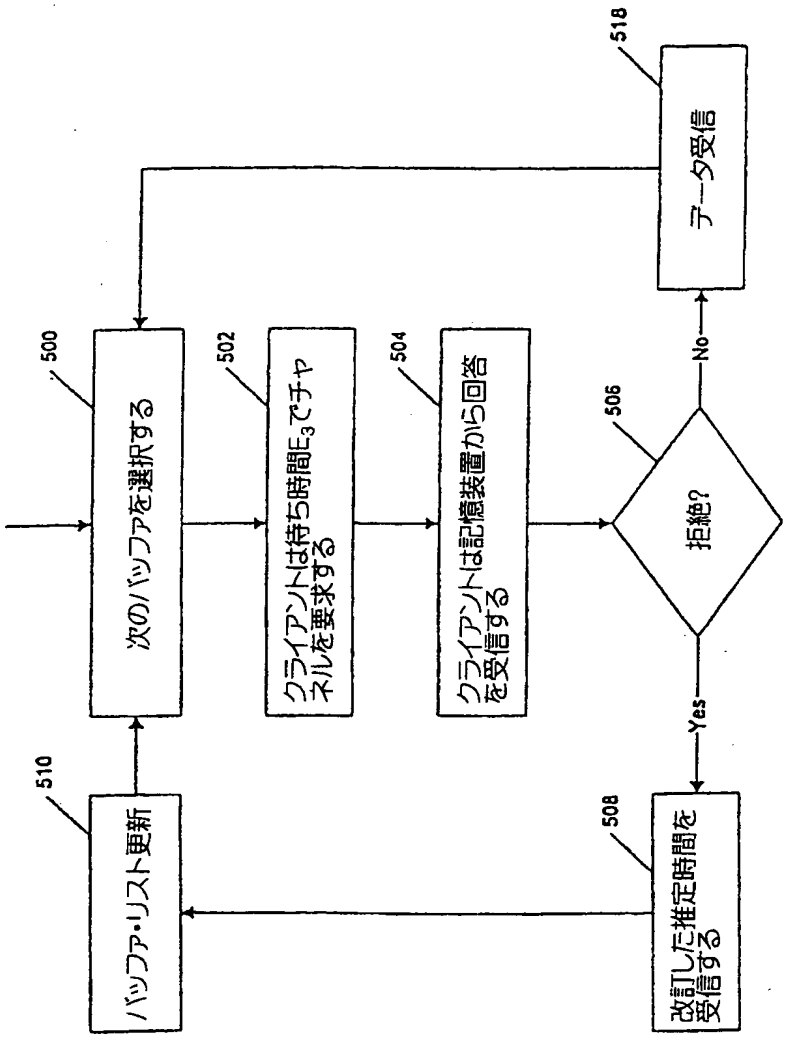
【補正対象書類名】図面

【補正対象項目名】図２０

【補正方法】変更

【補正内容】

【図２０】



【手続補正２０】

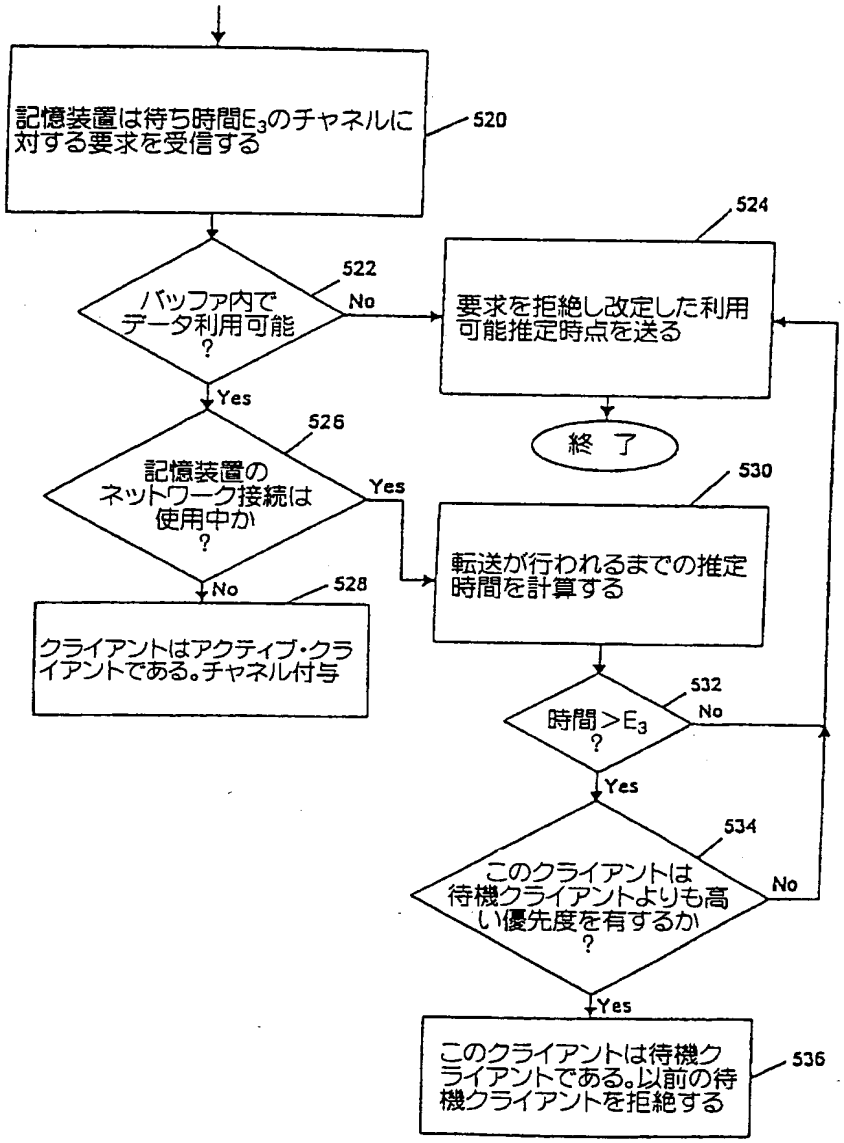
【補正対象書類名】図面

【補正対象項目名】図２１

【補正方法】変更

【補正内容】

【図２１】



【手続補正２１】

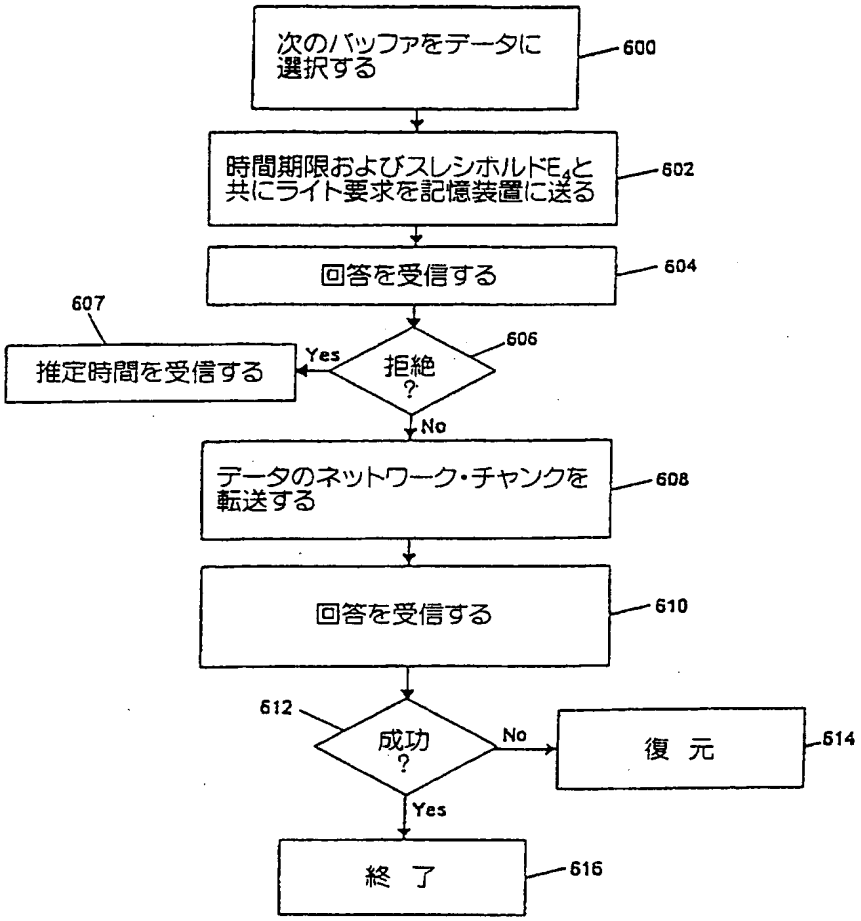
【補正対象書類名】図面

【補正対象項目名】図２２

【補正方法】変更

【補正内容】

【図２２】



【手続補正２２】

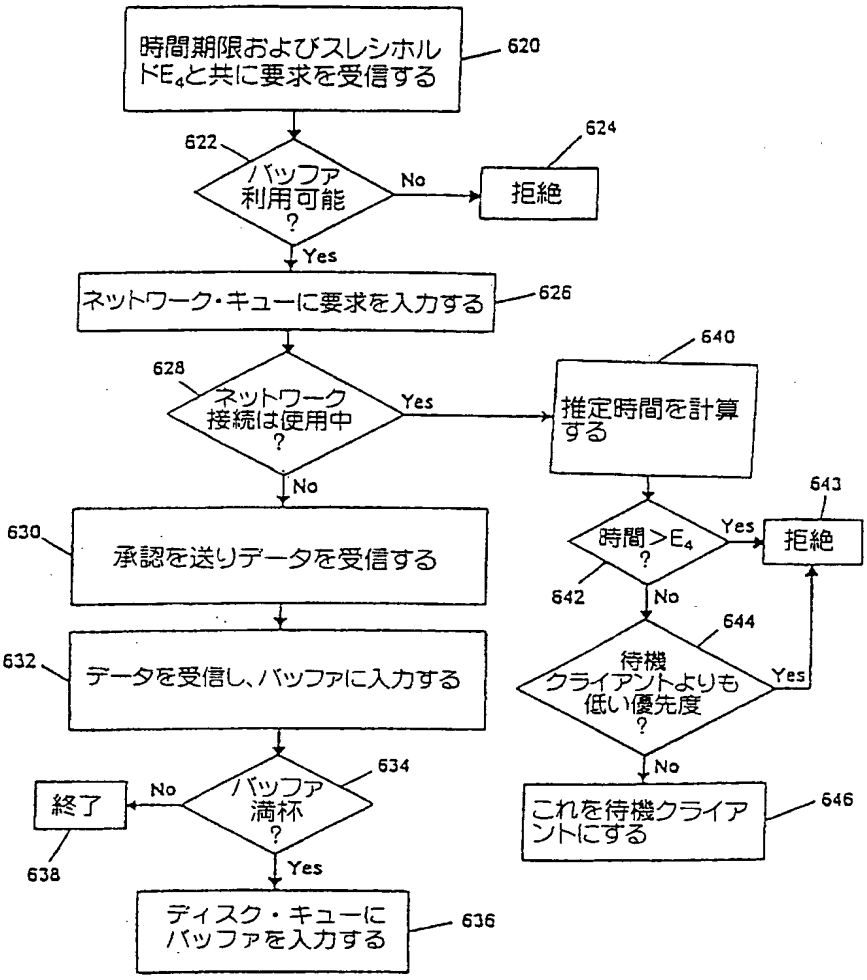
【補正対象書類名】図面

【補正対象項目名】図２３

【補正方法】変更

【補正内容】

【図２３】



【手続補正２３】

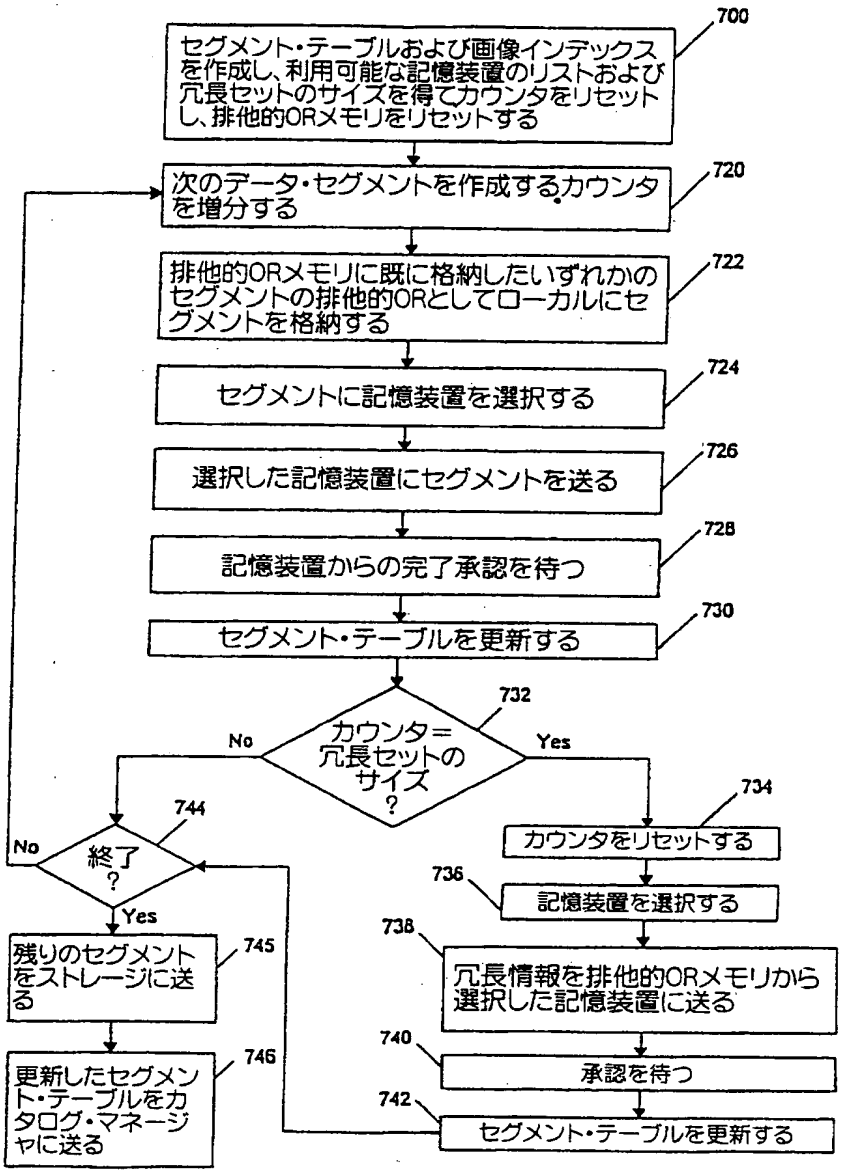
【補正対象書類名】図面

【補正対象項目名】図２４

【補正方法】変更

【補正内容】

【図２４】



【手続補正２４】

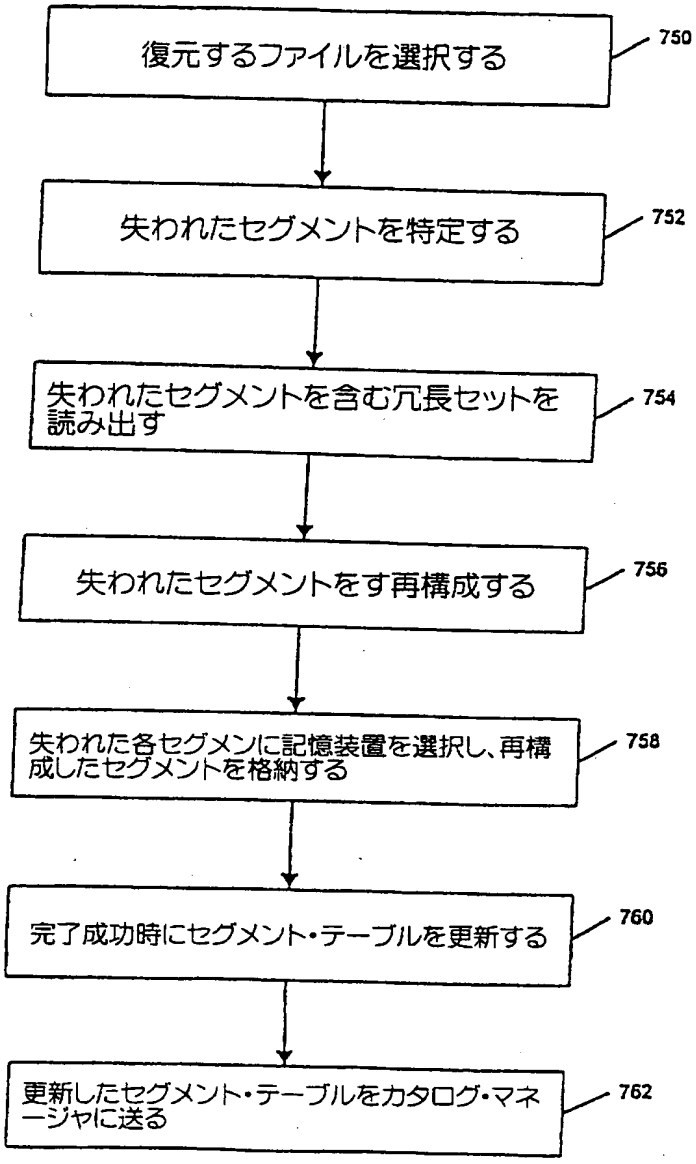
【補正対象書類名】図面

【補正対象項目名】図２５

【補正方法】変更

【補正内容】

【図２５】


復元するファイルを選択する
750
失われたセグメントを特定する
752
失われたセグメントを含む冗長セットを読み出す
754
失われたセグメントをす再構成する
756
失われた各セグメンに記憶装置を選択し、再構成したセグメントを格納する
758
完了成功時にセグメント・テーブルを更新する
760
更新したセグメント・テーブルをカタログ・マネージャに送る
762

【国際調査報告】

INTERNATIONAL SEARCH REPORT

Inter onal Application No
PCT/US 98/27199

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F11/20 G06F11/10 H04N7/173

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	R. TEWARI ET AL.: "High Availability in Clustered Multimedia Servers" PROC. 12TH INT. CONF. ON DATA ENGINEERING, 26 February 1996, pages 645-654, XP000632617 new orleans, la, usa see the whole document ---	1-10, 25-37
X	EP 0 701 198 A (STARLIGHT NETWORKS) 13 March 1996	1,5-9
Y	see column 7, line 43 - column 8, line 40 see column 27, line 18 - line 47 ---	16
	-/--	

[X] Further documents are listed in the continuation of box C. [X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 June 1999	23/06/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Absalom, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 98/27199

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	R. FLYNN ET AL.: "Disk Striping and Block Replication Algorithms for Video File Servers" PROC. INT. CONF. ON MULTIMEDIA COMPUTING AND SYSTEMS,17 June 1996, pages 590-597, XP002105211 Hiroshima, Japan see the whole document	1-6
Y	EP 0 740 247 A (HEWLETT-PACKARD COMPANY) 30 October 1996	16
A	see abstract	10-15
A	GB 2 299 424 A (MITSUBISHI DENKI KABUSHIKI KAISHA) 2 October 1996 see the whole document	1-37
A	B. NARENDRAN: "Data Distribution Algorithms for Load Balanced Fault-Tolerant Web Access" PROC. 16TH SYMP. ON RELIABLE DISTRIBUTED SYSTEMS,22 October 1997, pages 97-106, XP002105212 durham, nc, usa see the whole document	1-37
A	S. GHANDEHARIZADEH ET AL.: "Continuous Retrieval of Multimedia Data Using Parallelism" IEEE TRANS. ON KNOWLEDGA AND DATA ENGINEERING, vol. 5, no. 4, August 1993, pages 658-669, XP002105213 usa see the whole document	1-37
A	US 5 559 764 A (CHEN ET AL.) 24 September 1996 cited in the application	

1

Form PCT/ISA/210 (continuation of second sheet)(July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int... ...ional Application No
PCT/US 98/27199

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 701198	A	13-03-1996	US 5732239	A	24-03-1998
EP 740247	A	30-10-1996	US 5592612	A	07-01-1997
			JP 9026854	A	28-01-1997
GB 2299424	A	02-10-1996	JP 8329021	A	13-12-1996
			US 5630007	A	13-05-1997
US 5559764	A	24-09-1996	EP 0697660	A	21-02-1996
			JP 8069360	A	12-03-1996

Form PCT/ISA/210 (patent family annex) (July 1992)

フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	テーマコード(参考)
G06F 12/16	320	G06F 12/16	320L 5C064
H04N 7/173	610	H04N 7/173	610A

(31)優先権主張番号 09／054，761
(32)優先日 平成10年４月３日(1998. 4. 3)
(33)優先権主張国 米国（US）
(81)指定国 EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), AU, CA, CN, DE, GB, JP
(71)出願人 Metropolitan Technology Park, One Park West, Tewksbury, Massachusetts 01876, United States of America
(72)発明者 ラビノウィッツ，スタンレー
アメリカ合衆国マサチューセッツ州01886，ウエストフォード，ヴァイン・ブルック・ロード 12
(72)発明者 ジェイコブス，ハーバート・アール
アメリカ合衆国ニューハンプシャー州03051，ハドソン，サンライズ・ドライブ 17
(72)発明者 ジレット，リチャード・ベイカー，ジュニア
アメリカ合衆国マサチューセッツ州01886，ウエストフォード，プレザベーション・ウェイ 30
(72)発明者 ファッシアノ，ピーター・ジェイ
アメリカ合衆国マサチューセッツ州01760，ネイティック，コーチマン・レイン 30
Fターム(参考) 5B001 AA00 AB01 AC01 AD03
5B014 EA04 EB04 FB03 FB04 GC07
GD23 GD32 GD33 HA09
5B018 GA01 HA11 MA11 MA15
5B065 BA01 EA03 EA12 EA19 EA24
EA31 EA35 ZA08 ZA15
5B082 DE05 HA01
5C064 BA07 BB06 BC18 BD02 BD13

【要約の続き】
ての記憶装置上で均衡化する。この技法の組み合わせの結果、多数のアプリケーションおよび多数の記憶装置間で双方向にスケーラブルに多数の独立した高帯域データ・ストリームを転送可能なシステムが得られる。